Stochastic FISTA Step Search Algorithm for Convex Optimization

author names withheld

Under Review for OPT 2023

Abstract

In this paper, we propose an accelerated stochastic step search algorithm which combines an accelerated method with a fully adaptive step size parameter for convex problems in [18] with stochastic step search analysis in [15]. Under appropriate conditions on the accuracy of the estimates of gradient and function value our algorithm achieves expected iteration complexity of $\mathcal{O}(1/\sqrt{\epsilon})$ to reach an ϵ-accurate solution which satisfies $f(x) - f_* \leq \epsilon$. This complexity matches with the iteration complexity of deterministic Nesterov's accelerated and FISTA algorithms [1, 12, 13]. This paper continues the line of work on stochastic adaptive algorithms studied in [2, 3, 15] and is the first one to develop an accelerated gradient descent type algorithm in this domain.

1. Introduction

We consider the stochastic optimization problem $\min_{x\in\mathbb{R}^n} f(x)$, where exact values of the function and its gradient may not be computed exactly. Instead we assume that an optimization algorithm has access to some stochastic oracles which return estimations of the gradient values and the function values. We will specify these oracles and their properties later in the paper.

We make the following standard assumption.

Assumption 1 *We assume that f is convex and L-smooth.*

Stochastic Gradient Descent (SGD) [17] and its variants [7, 9] is one of the most popular approach for stochastic optimization. Most of the variants of SGD take a (usually diminishing) sequence of steps with precomputed step sizes along the direction of negative stochastic gradients. No function values are computed to check the progress of the algorithm. In deterministic setting a fixed step size whose value is usually proportional to the inverse of L guarantees algorithmic progress. In practical deterministic optimization, however, it is very common to try out step size parameter values and select them based on the achieved progress rather than a guess of a global constant L, which can be too conservative. Three of such adaptive algorithmic frameworks are line-search, trust-region and adaptive cubic regularization methods [6, 14]. Recently, all these methods have been analyzed in stochastic setting and shown to maintain favorable convergence properties under various conditions on stochastic function oracles [2, 3, 15].

Accelerated gradient methods, on the other hand, do not easily lend themselves to fully adaptive modes. The step size parameter can usually be decreased on any given iteration, but it cannot be increased between iterations as is done by the line-search, trust-region and adaptive cubic regularization methods. In addition the complexity analysis of accelerated first order methods is quite different from those of line-search, trust-region and adaptive cubic regularization methods. For the latter, complexity analysis relies on the fact that the objective function decreases by at least some fixed amount on each step. This, however, does not necessarily hold for accelerated methods.

Since the success of accelerated gradient methods, researchers have worked towards adapting this technique to stochastic settings [8, 10, 19–21]. Without assuming vanishing variance, SGD has a worse iteration complexity than Gradient Descent, which is $\mathcal{O}(1/\epsilon^2)$ compared to $\mathcal{O}(1/\epsilon)$ [11]. Naturally, applying the stochastic gradients straightforwardly to the accelerated momentum does not necessary lead to better complexity.

Contributions. In this paper we analyze an accelerated method, introduced in [18], which allows for increase and decrease of the step size parameter based on progress. We apply this algorithm in stochastic setting and extend the framework of [2, 15] to derive expected complexity of this method. We derive expected complexity bounds of $\mathcal{O}(1/\sqrt{\epsilon})$ Our analysis does not require gradient estimates to be unbiased, but it does require that the variance diminishes sufficiently fast. The gradient error is of a similar order as in [20], where an accelerated method is analyzed under deterministicly inexact setting and without lien search. In contrast with that paper our assumption on the gradient error is only in expectation. We are able to adapt the martingale arguments used in the analysis of prior adaptive stochastic methods [2, 3, 15] to the new setting of accelerated methods, were function decrease is no longer guaranteed on each iteration.

2. Function oracles

Let us define the oracles which our algorithm will call to estimate the values of our function and its gradient.

(i) **Stochastic first-order oracle (SFO(κ_g, δ_1)).** Given a point x, the oracle computes $g(x, \xi')$, a (random) estimate of the gradient $\nabla f(x)$, such that

$$\mathbb{P}_{\xi'}\left[\|g(x,\xi') - \nabla f(x)\| \leq \kappa_g \|\nabla f(x)\|\right] \geq 1 - \delta_1 \tag{1}$$

In summary, the input to the oracle is x , the output is $g(x, \xi')$, and the values (κ_g, δ_1) are intrinsic to the oracle. (we omit the dependence on ξ' for brevity)

(ii) **Stochastic zeroth-order oracle (SZO($\kappa_f, \alpha, \|g\|, \delta_2$)).** Given a point x and a constant $\alpha > 0$, the oracle computes $\tilde{f}(x, \xi)$, a (random) estimate of the function value $f(x)$. ξ is a random variable (whose distribution may depend on x), that satisfies

$$\mathbb{P}_{\xi}\left[|\tilde{f}(x,\xi) - f(x)| \leq \kappa_f \alpha^2 \|g\|^2\right] \geq 1 - \delta_2 \tag{2}$$

(we omit the dependence on ξ for brevity).

3. Algorithm Framework

In this section, we describe the accelerated stochastic step search algorithm, which is based on the variant of FISTA algorithm [1] introduced in [18] for deterministic exact optimization.

Step $(t^0_{k+1}, y_{k+1}) = \text{FISTAStep}(x_k, x_k^{prev}, t_k, \theta_k)$ is defined as follows.

$$t^0_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_k t_k^2}\right), \text{ and } y_{k+1} = x_k + \frac{t_k - 1}{t^0_{k+1}}(x_k - x_k^{prev}). \tag{3}$$

Algorithm 1 Stochastic FISTA Step Search Algorithm

```
Initialization:
Choose an initial point $x_0 \in \mathbb{R}^d$. Set $t_1 = 1, t_0 = 0$, let $x_0^{prev} = x_0$.
Set $0 < \gamma < 1, \theta_0 = \gamma$ with $\alpha_1 > 0$. Choose $\eta \in [1/2, 1]$.
for $k = 1, 2, \cdots,$ do
    Compute $(t_k^0, y_k) =$ FISTAStep$(x_{k-1}, x_{k-1}^{prev}, t_{k-1}, \theta_{k-1})$
    Compute the gradient estimate $g_k = g(y_k, \xi_k')$ using the $\kappa_g$ probabilistic first-order oracle
    and the reference point $p_{\alpha_k}(y_k) = y_k - \alpha_k g_k$.
    Compute the function estimates $f_k^y = \tilde{f}(y_k, \xi_k)$ and $f_k^p = \tilde{f}(p_{\alpha_k}(y_k), \xi_k)$.
    Check sufficient decrease condition $f_k^p \le f_k^y - \alpha_k \eta \|g_k\|^2$.
        If unsuccessful:
            Set $\left(x_k, x_k^{prev}\right) = \left(x_{k-1}, x_{k-1}^{prev}\right)$ and $t_k = t_{k-1}$, then set $\alpha_{k+1} = \gamma \alpha_k$ and $\theta_k = \frac{\theta_{k-1}}{\gamma}$.
        If successful:
            Set $\left(x_k, x_k^{prev}\right) = (p_{\alpha_k}(y_k), x_{k-1})$ and $t_k = t_k^0$, then set $\alpha_{k+1} = \frac{\alpha_k}{\gamma}$, and $\theta_k = \frac{\alpha_k}{\alpha_{k+1}}$.
end for
```

This update is slightly different from the one in [18]. The use of parameter θ_k allows for flexible step size selection. At each iteration, Algorithm 1 evaluates the computed gradient for the Armijo-type descent condition based on inexact function estimates. If the step is accepted, we call such iteration *successful iterations*.

Filtration. Algorithm 1 generates a random process given by the sequence $\{G_k, \mathcal{A}_k, F_k^y, F_k^p\}$. The random gradient estimate is denoted by G_k and its realization by $g_k = G_k(\omega)$, $\{F_k^y, F_k^p\}$ denotes estimates of $f(Y_k)$ and $f(p_{\mathcal{A}_k}(Y_k))$, similarly, we have $\alpha_k = \mathcal{A}_k(\omega)$ (step sizes). To formalize the conditioning on the past, let $\mathcal{F}_{k-1}^{G \cdot F}$ denote the σ-algebra generated by the random variables $G_0, \ldots, G_{k-1}$ and $F_0^y, F_0^p, \ldots, F_{k-1}^y, F_{k-1}^p$. Therefore $\{\mathcal{F}_{k-1}^{G \cdot F}\}_{k \ge 1}$ is a filtration.

Epsilon accurate solution and the hitting time N_ϵ to reach ϵ accuracy. Let x^* be a global minimizer of f and $f_* > -\infty$ is the minimum value of f over its domain. We say that x is an ϵ-accurate solution of problem $\min_{x \in \mathbb{R}^n} f(x)$ if it satisfies $f(x) - f_* \le \epsilon$. For a level of accuracy ϵ, we aim to derive a bound on the expected number of iterations the algorithm needs to reach the given accuracy level. Thus let us define N_ϵ as the hitting time for the event $\{f(X_k) - f_* \le \epsilon\}$, where k is the iteration of the algorithm.

Definition 1 (True iteration) *We say iteration k is* true *if $\|g_k - \nabla f(y_k)\| \le \kappa_g \|\nabla f(y_k)\|$ and $|f_k^y - f(y_k)| \le \kappa_f \alpha_k^2 \|g_k\|^2$, $|f_k^p - f(p_{\alpha_k}(y_k))| \le \kappa_f \alpha_k^2 \|g_k\|^2$, respectively. We use an indicator variable I_k^{true} to indicate that iteration k is true.*

The condition on the accuracy of the gradient in a true iteration is known as the norm condition and appears in various prior works [2, 4, 16] and similarly for function estimator [15]. Meanwhile, [22] assumes similar gradient noise condition in expectation, and as a result, their analysis bound the number of iterations until the expected function gap is small. Our assumption is more flexible than that in [2, 15] as we do not require κ_f to be smaller than a threshold. With these definitions, we present our key Assumption 2.

Assumption 2

(i) For every iteration k, the following bound holds:

$$\mathbb{E}\left[\|(\nabla f(y_k) - g_k)\|^2 | \mathcal{F}_{k-1}^{G\cdot F}\right] \leq \frac{1}{\alpha_k^2 t_k^2 \cdot k^{2+\beta}} \tag{4}$$

$$\mathbb{E}[(f_k^y - f(y_k))|\mathcal{F}_{k-1}^{G\cdot F}] \leq \frac{1}{\alpha_k t_k^2 \cdot k^{1+\beta'}}; \quad \mathbb{E}[(f(x_k) - f_k^p)|\mathcal{F}_{k-1}^{G\cdot F}] \leq \frac{1}{\alpha_k t_k^2 \cdot k^{1+\beta'}} \tag{5}$$

for some $\beta > 0$ and $\beta' > 0$.

(ii) Let $\eta \in [1/2, 1]$ and $\kappa_g \leq \frac{1-\eta}{2-\eta} \leq \frac{1}{3}$. There exists a probability $p \in (0.5, 1]$ such that:

$$\mathbb{P}(I_k^{true}|\mathcal{F}_{k-1}^{G\cdot F}) \geq p \textit{ for all } k \geq 0 \tag{6}$$

Assumption 2(i) implies that variance of gradient noise has to decrease as a particular rate. While this rate may appear complicated due dependence on α_k and t_k, α_k is bounded from above and we will show later in the paper that $t_k \sim k$. This means that the imposed rate of the decrease of the error is the same as that for the deterministic gradient error [20]. It follows from that work that accelerated convergence rate is lost if the error in the gradient does not decrease at this rate. Here we relax the bound on deterministic error and impose it only in expectation.

Assumption 2(ii) is used in prior work [2, 5, 15] and is needed to ensure that successful iterations happen sufficiently often when α_k is small enough (this is proven in Lemma 2 below). Satisfying these conditions depends on the nature of gradient and function estimates and is discussed, for example in [5, 15].

4. Analysis

4.1. Bound on the Expected Hitting Time

Now we are ready to present our analysis, starting with a property for the step sizes $\{\mathcal{A}_k\}_{k\geq 1}$.

Lemma 2 *Suppose that iteration k is true. Let $\kappa_g \leq \frac{1-\eta}{2-\eta} \leq \frac{1}{3}$. If $\alpha_k \leq \bar{\alpha} = \frac{1-\eta-(2-\eta)\kappa_g}{(0.5L+2\kappa_f)(1-\kappa_g)}$ then the k-th step is successful. In particular, this means $f_k^p \leq f_k^y - \eta\alpha_k \|g_k\|^2$.*

Without loss of generality, we assume that $\alpha_1 \geq \bar{\alpha}$. Lemma 2 states that good model estimates and sufficiently small step size leads to successful step. By Assumption 2(ii), we have that whenever the stepsize is below a threshold $\bar{\alpha}$, the iteration has a probability p to be successful. The first part of our analysis relies on this property to bound the expectation of the hitting time $\mathbb{E}(N_\epsilon)$ using the number of true successful iterations that are above a threshold $\bar{\alpha}$.

Theorem 3 (Bounding $\mathbb{E}(N_\epsilon)$ based on $\mathbb{E}(N_{TS})$) *We assume Assumption 2(ii). We consider N_{TS} the number of true successful iterations with $\alpha_k \geq \bar{\alpha}$:*

$N_{TS} = \sum_{k=1}^{N_\epsilon - 1} \mathbb{1}\{\mathcal{A}_k \geq \bar{\alpha}\} \cdot \mathbb{1}\{\text{Iteration } k \text{ is true and successful}\}$. *We have*

$$\mathbb{E}(N_\epsilon) \leq \frac{2p}{(2p-1)^2}\left(2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{7}$$

The main proof of Theorem 3 is heavily based on a prior framework for line search methods [5], where the same principle is further applied in [2, 15]. The remaining problem of analyzing Algorithm 1 focuses on the expectation of N_{TS}, which we present in the next section.

4.2. Bound on the Number of Large True Successful Iterations

In order to bound $\mathbb{E}(N_{TS})$, we consider the stochastic process $\mathcal{N}_k = 2\alpha_k^{succ} t_k^2 v_k + \|u_k\|^2$, where $v_k = f(x_k) - f_*$ and $u_k = (t_k - 1)x_k^{prev} + x^* - t_k x_k$. While v_k denotes the usual optimality gap, u_k resembles the usual term in deterministic FISTA algorithm analysis [1, 18]. We also denote the step size $\alpha_k^{succ} = \alpha_{k'}$, where $k' = \max\{k'' \text{ successful }, k'' \leq k\}$, the step size at the previous successful iteration prior to k. The next Lemma states that $\mathcal{N}_k$ do not change during unsuccessful iterations.

Lemma 4 (Equation for an unsuccessful step) *Let $\{x_k\}_{k\geq 1}$ be the iterates of Algorithm 1 and the stochastic process $\mathcal{N}_k$ be defined above. If iteration k is unsuccessful, then we have $\mathcal{N}_k = \mathcal{N}_{k-1}$.*

Lemma 5 shows how the quantity $\mathcal{N}_k$ changes in the successful steps of Algorithm 1, under the assumptions of convexity and L-Lipschitz smoothness of the objective function.

Lemma 5 (Bounds for a successful step) *Let $\{x_k\}_{k\geq 1}$ be the iterates of Algorithm 1. We assume Assumptions 1 and let $\eta \geq \frac{1}{2}$. If k is a successful iteration then we have $\mathcal{N}_k - \mathcal{N}_{k-1} \leq B_k$, where $B_k = 2\alpha_k t_k(\nabla f(y_k) - g_k)^\top u_{k-1} + 2\alpha_k t_k^2(f_k^y - f(y_k) + f(x_k) - f_k^p)$.*

In Lemma 5 B_k dictates how the noise of the model at iteration k affects the bound. Using Lemma 5 and the Assumption 2(i) we prove the following Theorem:

Theorem 6 (Upper bound of the number of large true successful steps) *Let $\{x_k\}_{k\geq 1}$ be the iterates of Algorithm 1. We assume Assumption 1 and Assumption 2(i). Then we have*

$$\mathbb{E}[N_{TS}] \leq \frac{\sqrt{2D_1}}{\sqrt{\bar{\alpha}\epsilon}}. \tag{8}$$

where $D_1 = 3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + \dfrac{3(\beta+2)^2}{\beta^2} + \dfrac{2(\beta+2)}{\beta+1} + \dfrac{6(\beta'+1)}{\beta'}$.

Our final result is a straightforward corollary of Theorem 3 and 6.

Theorem 7 *Let $\{x_k\}_{k\geq 1}$ be the iterates of Algorithm 1. We assume Assumption 1 and Assumption 2. The number of iterations N_ϵ to reach an ϵ accuracy solution satisfies the following bound:*

$$\mathbb{E}(N_\epsilon) \leq \frac{2p}{(2p-1)^2}\left(\frac{2\sqrt{2D_1}}{\sqrt{\bar{\alpha}\epsilon}} + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{9}$$

where $\bar{\alpha}$ is defined in Lemma 2 and D_1 is defined in Theorem 6.

The (expected) iteration complexity of Algorithm 1 is $\mathcal{O}(\epsilon^{-1/2})$, which matches with the iteration complexity of the deterministic accelerated momentum. Our result is especially helpful in the settings where κ_f is large, convergence is still guaranteed.

Remark 1 (Discussion and Future Work) *The recursive bound in Lemma 5 demonstrate that similar arguments can be used to analyze different stochastic algorithms where the deterministic analysis uses the telescoping sum. For example, ones can apply this framework to the stochastic gradient descent method and obtain the complexity of $\mathcal{O}(\epsilon)$ for convex setting. We can also derived similar bounds to prior reference [15] for non-convex and strongly convex setting. One possible question to explore is determining the circumstances under which our framework and Algorithm 1 can be applied to the proximal problem with an additional non-smooth term. With the lack of sufficient decrease in small proximal steps, such analysis is not straightforward. Thus this presents an interesting problem for future work.*

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

[2] A. S. Berahas, L. Cao, and K. Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31(2):1489–1518, 2021. doi: 10.1137/19M1291832. URL https://doi.org/10.1137/19M1291832.

[3] Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1:92–119, 04 2019. doi: 10.1287/ijoo.2019.0016.

[4] Richard G. Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28(1):251–265, 1991. doi: 10.1137/0728014. URL https://doi.org/10.1137/0728014.

[5] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018. doi: 10.1007/s10107-017-1137-4. URL https://doi.org/10.1007/s10107-017-1137-4.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programing*, 127(2):245–295, 2011.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[8] Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/ec5aa0b7846082a2415f0902f0da88f2-Paper.pdf.

[9] D. P. Kingma and J. Ba. ADAM: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, abs/1412.6980, 2014.

[10] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

[11] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2009.

[12] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. Translated as Soviet Math. Dokl.

[13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

[14] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

[15] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020. doi: 10.1137/18M1216250. URL https://doi.org/10.1137/18M1216250.

[16] Boris Polyak. Introduction to optimization, 07 2020.

[17] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[18] Katya Scheinberg, Donald Goldfarb, and Xi Bai. Fast First-Order Methods for Composite Convex Optimization with Backtracking. *Foundations of Computational Mathematics*, 14(3): 389–417, June 2014. ISSN 1615-3383. doi: 10.1007/s10208-014-9189-9. URL https://doi.org/10.1007/s10208-014-9189-9.

[19] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition, 2013.

[20] Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *CoRR*, abs/1109.2415, 2011. URL http://arxiv.org/abs/1109.2415.

[21] Trang H Tran, Katya Scheinberg, and Lam M Nguyen. Nesterov accelerated shuffling gradient method for convex optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21703–21732. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tran22a.html.

[22] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/vaswani19a.html.

Appendix A. Detail Proofs for Theorem 3

By the update rule of α_k in Algorithm 1, the stochastic process $\{\mathcal{A}_k\}_{k\geq 1}$ follows:

$$\mathcal{A}_1 = \alpha_1, \quad \mathcal{A}_{k+1} = \begin{cases} \gamma^{-1}\mathcal{A}_k & \text{if } I_k^{\text{succ}} = 1 \\ \gamma\mathcal{A}_k & \text{if } I_k^{\text{succ}} = 0. \end{cases} \tag{10}$$

Let us now define

$$I_k^{\text{large}} = \mathbb{1}\{\mathcal{A}_k > \bar{\alpha}\}, \tag{11}$$

and

$$I_k^{\text{succ}} = \mathbb{1}\{\text{Iteration } k \text{ is successful}\}. \tag{12}$$

Lemma 8 (Lemma 2.1 from [5]) *We assume Assumption 2(ii). Let N_ϵ be a hitting time for $\{\mathcal{A}_k\}$. Let W_k be a nonnegative stochastic process such that $\sigma(W_k) \subset \mathcal{F}_{k-1}$, for any $k \geq 1$. Then*

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k I_k^{\textit{true}}\right) \geq p\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k\right). \tag{13}$$

Similarly,

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k(1 - I_k^{\textit{true}})\right) \leq (1-p)\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k\right). \tag{14}$$

Proof The proof is a consequence of properties of expectation:

$$\mathbb{E}(I_k^{\text{true}} |\, W_k) = \mathbb{E}(\mathbb{E}(I_k^{\text{true}} |\, \mathcal{F}_{k-1}) |\, W_k) \geq \mathbb{E}(p |\, W_k) = p, \tag{15}$$

where we use the assumption that $\sigma(W_k) \subset \mathcal{F}_{k-1}$.

Hence by the law of total expectation, we have

$$\mathbb{E}(W_k I_k^{\text{true}}) = \mathbb{E}(W_k \mathbb{E}(I_k^{\text{true}} | W_k)) \geq p\mathbb{E}(W_k). \tag{16}$$

Similarly, we can derive

$$\mathbb{E}(\mathbb{1}\{k < N_\epsilon\} W_k I_k^{\text{true}}) \geq p\mathbb{E}(\mathbb{1}\{k < N_\epsilon\} W_k), \tag{17}$$

since $\mathbb{1}\{k < N_\epsilon\}$ is also determined by $\mathcal{F}_{k-1}$. Finally,

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k I_k^{\text{true}}\right) = \mathbb{E}\left(\sum_{k=1}^{\infty} \mathbb{1}\{k < N_\epsilon\} W_k I_k^{\text{true}}\right) \geq p\mathbb{E}\left(\sum_{k=1}^{\infty} \mathbb{1}\{k < N_\epsilon\} W_k\right) = p\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} W_k\right). \tag{18}$$

The second inequality is proved analogously.

A.1. Bounding the number of steps for which $\alpha_k \leq \bar{\alpha}$

In this subsection we derive a bound on $\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}})\right)$. The bound for $\mathbb{E}(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}})$ is in the next section.

Lemma 9 (Lemma 2.2 from [5]) *Let $\{\mathcal{A}_k\}_{k\geq 1}$ be the stochastic process defined in equation (10) and $\alpha_1 \geq \bar{\alpha}$. For any $l \in \{1, \ldots, T-1\}$, we have*

$$\sum_{k=1}^{l}(1 - I_k^{large}) I_k^{succ} \leq \frac{l}{2} \tag{19}$$

Proof We count the number of successful iteration that has $\mathcal{A}_k \leq \bar{\alpha}$. In this case $\mathcal{A}_{k+1} = \gamma^{-1}\mathcal{A}_k$. Since $\alpha_1 \geq \bar{\alpha}$, for each successful small iteration, when $\mathcal{A}_k$ gets increased by a factor of γ^{-1}, there has to be at least one iteration when $\mathcal{A}_k$ is decreased by the same factor. It follows that amongst all iterations, at most half can be successful and have $\mathcal{A}_k \leq \bar{\alpha}$.

Lemma 10 (Lemma 2.3 from [5])

Let $\{\mathcal{A}_k\}_{k\geq 1}$ be the stochastic process defined in equation (10) and $\alpha_1 \geq \bar{\alpha}$. Also we assume Assumption 2(ii). Then we have

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{large})\right) \leq \frac{1}{2p}\mathbb{E}(N_\epsilon) \tag{20}$$

Proof By Lemma 8 applied to $W_k = 1 - I_k^{\text{large}}$ we have

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}}) I_k^{\text{true}}\right) \geq p\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}})\right). \tag{21}$$

From the fact that all true iterations are successful when $\alpha_k \leq \bar{\alpha}$,

$$\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}}) I_k^{\text{true}} \leq \sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}}) I_k^{\text{succ}}. \tag{22}$$

Finally, from Lemma 9

$$\sum_{k=1}^{N_\epsilon - 1}(1 - I_k^{\text{large}}) I_k^{\text{true}} \leq \frac{N_\epsilon - 1}{2} \leq \frac{N_\epsilon}{2}. \tag{23}$$

Taking expectations in (22) and (23) and combining with (21), we obtain the result of the lemma.

A.2. Bounding the expected number of steps for which $\alpha_k > \bar{\alpha}$

Now we consider the bound on $\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}}\right)$. We introduce the additional notation $I_k^{\text{large+}} = \mathbb{1}\{\mathcal{A}_k > \bar{\alpha}\} + \mathbb{1}\{\mathcal{A}_k = \bar{\alpha}\}$. In other words $I_k^{\text{large+}} = 1$ when either $I_k^{\text{large}} = 1$ or $\mathcal{A}_k = \bar{\alpha}$. We now define:

- $N_{FS} = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}}(1 - I_k^{\text{true}}) I_k^{\text{succ}}$, which is the number of false successful iterations with $\mathcal{A}_k \geq \bar{\alpha}$.
- $N_F = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}}(1 - I_k^{\text{true}})$, which is the number of false iterations with $\mathcal{A}_k \geq \bar{\alpha}$.
- $N_{TS} = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}} I_k^{\text{true}} I_k^{\text{succ}}$, which is the number of true successful iterations with $\mathcal{A}_k \geq \bar{\alpha}$.
- $N_T = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}} I_k^{\text{true}}$, which is the number of true iterations with $\mathcal{A}_k \geq \bar{\alpha}$.
- $N_{TU} = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}} I_k^{\text{true}}(1 - I_k^{\text{succ}})$, which is the number of true unsuccessful iterations with $\mathcal{A}_k > \bar{\alpha}$.
- $N_U = \sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}}(1 - I_k^{\text{succ}})$, which is the number of unsuccessful iterations with $\mathcal{A}_k > \bar{\alpha}$.

Since $\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}}\right) = \mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}}(1 - I_k^{\text{true}})\right) + \mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large}} I_k^{\text{true}}\right) \leq \mathbb{E}(N_F) + \mathbb{E}(N_T)$, our goal is to bound $\mathbb{E}(N_F) + \mathbb{E}(N_T)$.

Another key observation is that

$$N_T \leq N_{TS} + N_{TU} \leq N_{TS} + N_U, \tag{24}$$

where the first inequality follows from the fact that for all $k < T$ and for all realizations, $(I_k^{\text{large+}} - I_k^{\text{large}}) I_k^{\text{true}}(1 - I_k^{\text{succ}}) = 0$, in other words there are no true unsuccessful iterations when $\mathcal{A}_k = \bar{\alpha}$.

Lemma 11 (Lemma 2.5 from [5]) *Let $\{\mathcal{A}_k\}_{k \geq 1}$ be the stochastic process defined in equation (10) and $\alpha_1 \geq \bar{\alpha}$. For any $l \in \{1, \ldots, N_\epsilon - 1\}$ we have*

$$\sum_{k=0}^{l} I_k^{\textit{large}}(1 - I_k^{\textit{succ}}) \leq \sum_{k=1}^{l} I_k^{\textit{large+}} I_k^{\textit{succ}} + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right) \tag{25}$$

Proof $\mathcal{A}_k$ is increased on successful iterations and decreased on unsuccessful ones. Hence the total number of steps when $\mathcal{A}_k > \bar{\alpha}$ and $\mathcal{A}_k$ is decreased, is bounded by the total number of steps when $\mathcal{A}_k \geq \bar{\alpha}$ is increased plus the number of steps it is required to reduce $\mathcal{A}_k$ from its initial value α_1 to $\bar{\alpha}$.

From Lemma 11 applied to $l = N_\epsilon - 1$, we can deduce that

$$N_U \leq N_{FS} + N_{TS} + \log_\gamma(\bar{\alpha}/\alpha_1). \tag{26}$$

We also have the following lemma.

Lemma 12 (Lemma 2.6 from [5])

$$\mathbb{E}(N_F) \leq \frac{1-p}{p}\mathbb{E}(N_T). \tag{27}$$

Proof By applying both inequalities in Lemma 8 with $W_k = I_k^{\text{large+}}$, we obtain

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}} I_k^{\text{true}}\right) \geq p\mathbb{E}\left(\sum_{k=1}^{N_\epsilon - 1} I_k^{\text{large+}}\right) \tag{28}$$

and

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large+}}(1-I_k^{\text{true}})\right) \leq (1-p)\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large+}}\right) \tag{29}$$

which gives us

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large+}}(1-I_k^{\text{true}})\right) \leq \frac{1-p}{p}\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large+}} I_k^{\text{true}}\right). \tag{30}$$

Lemma 13 (Lemma 2.7 from [5]) *Under the condition that $p > 1/2$, we have*

$$\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{large}\right) \leq \frac{1}{2p-1}\left(2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{31}$$

Proof Recall that $\mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large}}\right) = \mathbb{E}(N_F + N_T)$. Using (24) and (27) it follows that

$$\mathbb{E}(N_{FS}) \leq \mathbb{E}(N_F) \leq \frac{1-p}{p}\mathbb{E}(N_T) \leq \frac{1-p}{p}\mathbb{E}(N_{TS}+N_U) = \frac{1-p}{p}[\mathbb{E}(N_{TS}) + \mathbb{E}(N_U)]. \tag{32}$$

Taking into account (26) we have

$$\mathbb{E}(N_U) \leq \mathbb{E}(N_{FS}) + \mathbb{E}(N_{TS}) + \log_\gamma(\bar{\alpha}/\alpha_1). \tag{33}$$

Plugging this into (32) we obtain

$$\mathbb{E}(N_{FS}) \leq \frac{1-p}{p}\left[\mathbb{E}(N_{FS}) + 2\mathbb{E}(N_{TS}) + \log_\gamma(\bar{\alpha}/\alpha_1)\right] \tag{34}$$

and, hence,

$$\frac{2p-1}{p}\mathbb{E}(N_{FS}) \leq \frac{1-p}{p}\left[2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right]. \tag{35}$$

This finally implies

$$\mathbb{E}(N_{FS}) \leq \frac{1-p}{2p-1}\left[2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right]. \tag{36}$$

Now we can bound the expected total number of iterations when $\alpha_k > \bar{\alpha}$, using (33) and (36) and adding the terms to obtain the result of the lemma, namely,

$$\mathbb{E}(N_F + N_T) \leq \mathbb{E}(N_F + N_U + N_{TS}) \leq \frac{1}{p}\mathbb{E}(N_U + N_{TS}) \leq \frac{1}{2p-1}\left(2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{37}$$

A.3. Final bound on the expected stopping time

We finally prove the main theorem for the Step Search Framework which follows from Lemmas 10 and 13.

Proof Clearly

$$\mathbb{E}(N_\epsilon) = \mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} I_k^{\text{large}}\right) + \mathbb{E}\left(\sum_{k=1}^{N_\epsilon-1} (1 - I_k^{\text{large}})\right) \tag{38}$$

and, hence, using Lemmas 10 and 13 we have

$$\mathbb{E}(N_\epsilon) \leq \frac{1}{2p}\mathbb{E}(N_\epsilon) + \frac{1}{2p-1}\left(2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{39}$$

The result of the theorem follows.

Appendix B. Detailed Proofs for FISTA update

Firstly, we present Lemma 14 showing how $\{\alpha_k, t_k\}$ behaves in the updates of Algorithm 1

B.1. Lemma 14 and Proof

Using similar arguments as [18], we can prove the following Lemma:

Lemma 14 *Let $\{\alpha_k, t_k\}$ be the parameters in Algorithm 1 and let S be the set of successful iterations. Then we have*

- $\alpha_{k-1}^{succ} t_{k-1}^2 \geq \alpha_k t_k (t_k - 1)$ *for every successful iteration k,*
- $\alpha_K t_K^2 \geq (\sum_{k \in S, k \leq K} \sqrt{\alpha_k}/2)^2$ *for any number of iterations K.*

Proof

- From the update of the algorithm and the fact that k is successful:

$$t_k = t_k^0 = \frac{1 + \sqrt{1 + 4\theta_{k-1} t_{k-1}^2}}{2}, \tag{40}$$

  we have $\theta_{k-1} t_{k-1}^2 = t_k(t_k - 1)$.

  The statement of the lemma $\alpha_{k-1}^{succ} t_{k-1}^2 \geq \alpha_k t_k (t_k - 1)$ is equivalent to

$$\alpha_{k-1}^{succ} \geq \alpha_k \theta_{k-1}. \tag{41}$$

  Let k' be the previous successful iteration prior to k ($k' < k$), i.e. $\alpha_{k-1}^{succ} = \alpha_{k'}$. By the definition of k' we have $\alpha_{k'+1} = \frac{\alpha_{k'}}{\gamma}$, and $\theta_{k'} = \gamma$. Since the iterations from $k' + 1$ to $k - 1$ are unsuccessful (if $k' < k - 1$):

$$\alpha_k = \alpha_{k'+1} \cdot \gamma^{k-k'-1} \tag{42}$$

$$\theta_{k-1} = \theta_{k'} \gamma^{k'+1-k} \tag{43}$$

thus

$$\alpha_k = \alpha_{k'+1} \cdot \gamma^{k-k'-1} = \alpha_{k'} \cdot \gamma^{k-k'-2} \tag{44}$$

$$\theta_{k-1} = \theta_{k'} \gamma^{k'+1-k} = \gamma^{k'+2-k} \tag{45}$$

and it proves the statement:

$$\alpha_k \theta_{k-1} = \alpha_{k'} \cdot \gamma^{k-k'-2} \cdot \gamma^{k'+2-k} = \alpha_{k'}. \tag{46}$$

- $\alpha_K t_K^2 \geq (\sum_{k \in S, k \leq K} \sqrt{\alpha_k}/2)^2$ for any number of iterations K. This follows from the proof of Lemma 3.4 in [18]

B.2. Proof of Lemma 2: Accurate gradient and function estimates and small step size ⇒ successful step

Proof Using the triangle inequality we get $\|g_k\| \geq (1-\kappa_g)\|\nabla f(y_k)\|$. The L-smoothness of f and the fact that $\|g_k - \nabla f(y_k)\| \leq \kappa_g \|\nabla f(y_k)\|$ yield

$$\begin{aligned} f(p_{\alpha_k}(y_k)) &\leq f(y_k) - \alpha_k(\nabla f(y_k) - g_k)^T g_k - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \\ &\leq f(y_k) + \kappa_g \alpha_k \|g_k\| \|\nabla f(y_k)\| - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \\ &\leq f(y_k) + \frac{\kappa_g}{1-\kappa_g}\alpha_k \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \end{aligned} \tag{47}$$

Since the estimates satisfies $|f_k^y - f(y_k)| \leq \kappa_f \alpha_k^2 \|g_k\|^2$, $|f_k^p - f(p_{\alpha_k}(y_k))| \leq \kappa_f \alpha_k^2 \|g_k\|^2$, we obtain

$$\begin{aligned} f_k^p &\leq \kappa_f \alpha_k^2 \|g_k\|^2 + f(p_{\alpha_k}(y_k)) \\ &\leq \kappa_f \alpha_k^2 \|g_k\|^2 + f(y_k) + \frac{\kappa_g}{1-\kappa_g}\alpha_k \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 + \kappa_f \alpha_k^2 \|g_k\|^2 \\ &\leq \kappa_f \alpha_k^2 \|g_k\|^2 + (f(y_k) - f_k^y) + f_k^y + \frac{\kappa_g}{1-\kappa_g}\alpha_k \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \\ &\leq f_k^y + 2\kappa_f \alpha_k^2 \|g_k\|^2 + \frac{\kappa_g}{1-\kappa_g}\alpha_k \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 . \end{aligned} \tag{48}$$

The result follows by noting $f_k^p \leq f_k^y - \alpha_k \|g_k\|^2 \left(\frac{1-2\kappa_g}{1-\kappa_g} - \alpha_k \left(\frac{L}{2} + 2\kappa_f\right)\right)$.

■

B.3. Proof of Lemma 4: Equation for an unsuccessful step

Proof From the updates of Algorithm 1, we get the following dynamics of the triplet $\left(x_k, x_k^{prev}, t_k\right)$:

$$\left(x_k, x_k^{prev}, t_k\right) = \left(x_{k-1}, x_{k-1}^{prev}, t_{k-1}\right) \text{ for every unsuccessful iteration } k \tag{49}$$

From this property and from the definition of u_k and v_k (involving t_k, x_k and x_k^{prev}), we get that

$$u_k = u_{k-1} \text{ and } v_k = v_{k-1} \tag{50}$$

Finally, from the definition of α_k^{succ}, it follows that when k is not successful then $\alpha_k^{succ} = \alpha_{k-1}^{succ}$. Combining these with the fact that t_k, u_k and v_k are invariant, we obtain the desired equation $\mathcal{N}_k = \mathcal{N}_{k-1}$ for every unsuccessful iteration k. ■

B.4. Proof of Lemma 5: Bounds for a successful step

Proof Since k is a successful iteration, we have $f_k^p \leq f_k^y - \alpha_k \eta \|g_k\|^2$. Also $(x_k, x_k^{prev}) = (p_{\alpha_k}(y_k), x_{k-1}); t_k = t_k^0$. Hence we have

$$f(x_k) \leq f(y_k) - \alpha_k \eta \|g_k\|^2 + [f(x_k) - f_k^p] + [f_k^y - f(y_k)]. \tag{51}$$

Combining this with the convexity of f at x and y_k that

$$f(x) \geq f(y_k) + \nabla f(y_k)^\top (x - y_k), \tag{52}$$

we have

$$f(x) - f(x_k) \geq \nabla f(y_k)^\top (x - y_k) + \alpha_k \eta \|g_k\|^2 - [(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \tag{53}$$

Apply this equation for $x = x_{k-1}$ and $x = x^*$ we have

$$\begin{aligned} f(x_{k-1}) - f(x_k) &\geq \nabla f(y_k)^\top (x_{k-1} - y_k) + \alpha_k \eta \|g_k\|^2 - [(f_k^y - f(y_k)) + (f(x_k) - f_k^p)], \\ f(x^*) - f(x_k) &\geq \nabla f(y_k)^\top (x^* - y) + \alpha_k \eta \|g_k\|^2 - [(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \end{aligned} \tag{54}$$

Let $v_k = f(x_k) - f(x^*)$ then

$$\begin{aligned} v_{k-1} - v_k &\geq \nabla f(y_k)^\top (x_{k-1} - y_k) + \alpha_k \eta \|g_k\|^2 - [(f_k^y - f(y_k)) + (f(x_k) - f_k^p)], \\ -v_k &\geq \nabla f(y_k)^\top (x^* - y_k) + \alpha_k \eta \|g_k\|^2 - [(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \end{aligned} \tag{55}$$

Multiplying the first equality by $t_k - 1$ and adding to the second equality, we get

$$\begin{aligned} (t_k - 1)v_{k-1} - t_k v_k \geq\ & \nabla f(y_k)^\top ((t_k - 1)x_{k-1} + x^* - t_k y_k) + t_k \alpha_k \eta \|g_k\|^2 \\ & - t_k[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \end{aligned} \tag{56}$$

Multiplying the previous inequality by t_k we have

$$\begin{aligned} t_k(t_k - 1)v_{k-1} - t_k^2 v_k \geq\ & t_k \nabla f(y_k)^\top ((t_k - 1)x_{k-1} + x^* - t_k y_k) + t_k^2 \alpha_k \eta \|g_k\|^2 \\ & - t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \end{aligned} \tag{57}$$

Now using the fact that $x_k = p_{\alpha_k}(y_k) = y_k - \alpha_k g_k$, we get that $\alpha_k g_k = y_k - x_k$. Therefore

$$\begin{aligned} & t_k(t_k - 1)v_{k-1} - t_k^2 v_k \\ & \geq t_k[\nabla f(y_k) - g_k + g_k]^\top ((t_k - 1)x_{k-1} + x^* - t_k y_k) + \frac{\eta}{\alpha_k} \|t_k(y_k - x_k)\|^2 \end{aligned}$$

$$
\begin{aligned}
&- t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)] \\
&\geq t_k g_k^\top((t_k-1)x_{k-1} + x^* - t_k y_k) + \frac{\eta}{\alpha_k}\left\|t_k(y_k - x_k)\right\|^2 \\
&+ t_k[\nabla f(y_k) - g_k]^\top((t_k-1)x_{k-1} + x^* - t_k y_k) - t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)] \\
&\geq t_k \frac{1}{\alpha_k}(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k y_k) + \frac{\eta}{\alpha_k}\left\|t_k(y_k - x_k)\right\|^2 \\
&+ t_k[\nabla f(y_k) - g_k]^\top((t_k-1)x_{k-1} + x^* - t_k y_k) - t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)] \\
&\geq t_k \frac{1}{\alpha_k}(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k x_k) + t_k \frac{1}{\alpha_k}(y_k - x_k)^\top t_k(x_k - y_k) \\
&+ \frac{\eta}{\alpha_k}\left\|t_k(y_k - x_k)\right\|^2 + t_k[\nabla f(y_k) - g_k]^\top((t_k-1)x_{k-1} + x^* - t_k y_k) \\
&- t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)] \\
&\geq t_k \frac{1}{\alpha_k}(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k x_k) - \frac{1}{\alpha_k}\left\|t_k(x_k - y_k)\right\|^2 \\
&+ \frac{\eta}{\alpha_k}\left\|t_k(y_k - x_k)\right\|^2 + t_k[\nabla f(y_k) - g_k]^\top((t_k-1)x_{k-1} + x^* - t_k y_k) \\
&- t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)].
\end{aligned} \tag{58}
$$

Multiply the final statement by $2\alpha_k$ we get

$$
\begin{aligned}
&2\alpha_k t_k(t_k-1)v_{k-1} - 2\alpha_k t_k^2 v_k \\
&\geq 2t_k(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k x_k) - 2\left\|t_k(x_k - y_k)\right\|^2 \\
&+ 2\eta\left\|t_k(y_k - x_k)\right\|^2 + 2\alpha_k t_k[\nabla f(y_k) - g_k]^\top((t_k-1)x_{k-1} + x^* - t_k y_k) \\
&- 2\alpha_k t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)].
\end{aligned} \tag{59}
$$

Now we consider the following term:

$$
A = 2t_k(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k x_k) - \left\|t_k(x_k - y_k)\right\|^2. \tag{60}
$$

Let $a = t_k(y_k - x_k), b = (t_k-1)x_{k-1} + x^* - t_k x_k$ we have

$$
\begin{aligned}
A &= 2t_k(y_k - x_k)^\top((t_k-1)x_{k-1} + x^* - t_k x_k) - \left\|t_k(x_k - y_k)\right\|^2 \\
&= 2ab - a^2 \\
&= b^2 - b^2 + 2ab - a^2 = b^2 - (a-b)^2 \\
&= \left\|(t_k-1)x_{k-1} + x^* - t_k x_k\right\|^2 - \left\|(t_k-1)x_{k-1} + x^* - t_k x_k - t_k(y_k - x_k)\right\|^2 \\
&= \left\|(t_k-1)x_{k-1} + x^* - t_k x_k\right\|^2 - \left\|(t_k-1)x_{k-1} + x^* - t_k y_k\right\|^2.
\end{aligned} \tag{61}
$$

Since k is a successful iteration, we have that $x_k^{prev} = x_{k-1}$. From the definition of u_k

$$
u_k = (t_k-1)x_k^{prev} + x^* - t_k x_k, \tag{62}
$$

we have that

$$
u_k = (t_k-1)x_{k-1} + x^* - t_k x_k. \tag{63}
$$

Now from the FISTA step:

$$y_k = x_{k-1} + \frac{t_{k-1}-1}{t_k}(x_{k-1} - x_{k-1}^{prev}), \tag{64}$$

we have

$$t_k y_k = t_k x_{k-1} + (t_{k-1}-1)(x_{k-1} - x_{k-1}^{prev}). \tag{65}$$

Combining these equations with (61) we get:

$$\begin{aligned} A &= \|(t_k-1)x_{k-1} + x^* - t_k x_k\|^2 - \|(t_k-1)x_{k-1} + x^* - t_k y_k\|^2 \\ &= \|u_k\|^2 - \left\|(t_k-1)x_{k-1} + x^* - t_k x_{k-1} - (t_{k-1}-1)(x_{k-1} - x_{k-1}^{prev})\right\|^2 \\ &= \|u_k\|^2 - \left\|-t_{k-1}x_{k-1} + x^* + (t_{k-1}-1)x_{k-1}^{prev}\right\|^2 \\ &= \|u_k\|^2 - \|-u_{k-1}\|^2 . \end{aligned} \tag{66}$$

Substituting back to previous equality (59) we get

$$\begin{aligned} &2\alpha_k t_k(t_k-1)v_{k-1} - 2\alpha_k t_k^2 v_k \\ &\geq \|u_k\|^2 - \|u_{k-1}\|^2 - \|t_k(x_k-y_k)\|^2 + 2\eta\,\|t_k(y_k-x_k)\|^2 + 2\alpha_k t_k[\nabla f(y_k) - g_k]^\top(-u_{k-1}) \\ &- 2\alpha_k t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)] \\ &\geq \|u_k\|^2 - \|u_{k-1}\|^2 + (2\eta-1)\,\|t_k(y_k-x_k)\|^2 - 2\alpha_k t_k[\nabla f(y_k) - g_k]^\top u_{k-1} \\ &- 2\alpha_k t_k^2[(f_k^y - f(y_k)) + (f(x_k) - f_k^p)]. \end{aligned} \tag{67}$$

Note that $\eta \geq \frac{1}{2}$ we have $(2\eta-1)\,\|t_k(y_k-x_k)\|^2 \geq 0$ and

$$2\alpha_k t_k(t_k-1)v_{k-1} - 2\alpha_k t_k^2 v_k \geq \|u_k\|^2 - \|u_{k-1}\|^2 - B_k \tag{68}$$

where $B_k = 2\alpha_k t_k(\nabla f(y_k) - g_k)^\top u_{k-1} + 2\alpha_k t_k^2(f_k^y - f(y_k) + f(x_k) - f_k^p)$.

From Lemma 14 we have $\alpha_{k-1}^{succ} t_{k-1}^2 \geq \alpha_k t_k(t_k-1)$ for every successful iteration k. Also since k is successful, we have $\alpha_k = \alpha_k^{succ}$. Thus we have the final bound of Lemma 5. ■

B.5. Proof of Theorem 6: Upper bound of the number of large true successful steps

Proof We have

$$\mathcal{N}_K - \mathcal{N}_0 = \sum_{k\in S, k\leq K} (\mathcal{N}_k - \mathcal{N}_{k-1}), \tag{69}$$

for every index K. We apply Lemma 5 for successful iterations and have

$$\mathcal{N}_K - \mathcal{N}_0 = \sum_{k\in S, k\leq K} (\mathcal{N}_k - \mathcal{N}_{k-1}) \leq \sum_{k\in S, k\leq K} B_k \tag{70}$$

By the definition of $\mathcal{N}_k$:

$$\left(2\alpha_K^{succ} t_K^2 v_K + \|u_K\|^2\right) - \left(2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2\right) \leq \sum_{k\in S, k\leq K} B_k \tag{71}$$

where by the definition of B_k

$$B_k = 2\alpha_k t_k(\nabla f(y_k) - g_k)^\top u_{k-1} + 2\alpha_k t_k^2(f_k^y - f(y_k) + f(x_k) - f_k^p). \tag{72}$$

We further have

$$\begin{aligned}\sum_{k\in S, k\leq K} B_k &\leq \left|\sum_{k\in S, k\leq K} B_k\right| \leq \sum_{k\in S, k\leq K} |B_k| \leq \sum_{k=1}^{K} |B_k| \\ &\leq \sum_{k=1}^{K} |2\alpha_k t_k(\nabla f(y_k) - g_k)^\top u_{k-1} + 2\alpha_k t_k^2(f_k^y - f(y_k) + f(x_k) - f_k^p)| \\ &\leq \sum_{k=1}^{K} 2\alpha_k t_k \|(\nabla f(y_k) - g_k)\| \cdot \|u_{k-1}\| + \sum_{k=1}^{K} 2\alpha_k t_k^2 |(f_k^y - f(y_k)) + (f(x_k) - f_k^p)|\end{aligned} \tag{73}$$

and plugging back to the previous inequality, we have the following for every iteration K:

$$\begin{aligned}2\alpha_K^{succ} t_K^2 v_K + \|u_K\|^2 \leq \left(2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2\right) + \sum_{k=1}^{K} 2\alpha_k t_k \|(\nabla f(y_k) - g_k)\| \cdot \|u_{k-1}\| \\ + \sum_{k=1}^{K} 2\alpha_k t_k^2 |(f_k^y - f(y_k)) + (f(x_k) - f_k^p)|\end{aligned} \tag{74}$$

Ignoring the positive term $2\alpha_K^{succ} t_K^2 v_K$, we have

$$\begin{aligned}\|u_K\|^2 \leq \left(2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2\right) + \sum_{k=1}^{K} 2\alpha_k t_k \|(\nabla f(y_k) - g_k)\| \cdot \|u_{k-1}\| \\ + \sum_{k=1}^{K} 2\alpha_k t_k^2 |(f_k^y - f(y_k)) + (f(x_k) - f_k^p)|\end{aligned} \tag{75}$$

Changing the index of the second term, we have

$$\begin{aligned}\|u_K\|^2 \leq &\left(2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2\right) + \sum_{k=0}^{K-1} 2\alpha_{k+1} t_{k+1} \|(\nabla f(y_{k+1}) - g_{k+1})\| \cdot \|u_k\| \\ &+ \sum_{k=0}^{K-1} 2\alpha_{k+1} t_{k+1}^2 |(f_{k+1}^y - f(y_{k+1})) + (f(x_{k+1}) - f_{k+1}^p)| \\ &\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + 2\alpha_1 t_1 \|(\nabla f(y_1) - g_1)\| \cdot \|u_0\| \\ &+ \sum_{k=0}^{K-1} 2\alpha_{k+1} t_{k+1}^2 |(f_{k+1}^y - f(y_{k+1})) + (f(x_{k+1}) - f_{k+1}^p)| \\ &+ \sum_{k=1}^{K} 2\alpha_{k+1} t_{k+1} \|(\nabla f(y_{k+1}) - g_{k+1})\| \cdot \|u_k\|\end{aligned}$$

$$\leq S_K + \sum_{k=1}^{K} \mu_{k+1} \|u_k\|, \tag{76}$$

where

$$\begin{aligned} S_K &= 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + 2\alpha_1 t_1 \|(\nabla f(y_1) - g_1)\| \|u_0\| + \\ &\quad + \sum_{k=0}^{K-1} 2\alpha_{k+1} t_{k+1}^2 |(f_{k+1}^y - f(y_{k+1})) + (f(x_{k+1}) - f_{k+1}^p)| \\ \mu_k &= 2\alpha_k t_k \|(\nabla f(y_k) - g_k)\| \end{aligned} \tag{77}$$

We have

$$S_k = 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + 2\alpha_1 t_1 \|(\nabla f(y_1) - g_1)\| \|u_0\| + F_k \tag{78}$$

where $F_k = \sum_{i=0}^{k-1} 2\alpha_{i+1} t_{i+1}^2 |(f_{i+1}^y - f(y_{i+1})) + (f(x_{i+1}) - f_{i+1}^p)|$.

$$\begin{aligned} (S_k)^{1/2} &= \sqrt{2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \mu_1 \|u_0\| + F_k} \\ &\leq \sqrt{2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \mu_1^2 + \|u_0\|^2 + F_k} \\ &\leq \sqrt{2\alpha_0^{succ} t_0^2 v_0 + 2\left\|u_0\right\|^2 + \mu_1^2 + F_k} \\ &\leq \left(2\alpha_0^{succ} t_0^2 v_0 + 2\left\|u_0\right\|^2\right)^{1/2} + \mu_1 + (F_k)^{1/2} = C_1 + \mu_1 + (F_k)^{1/2} \end{aligned} \tag{79}$$

where $C_1 = \left(2\alpha_0^{succ} t_0^2 v_0 + 2\left\|u_0\right\|^2\right)^{1/2}$ is a constant.

Thus applying Lemma 1 in [20], we have for every k:

$$\|u_k\| \leq \frac{1}{2} \sum_{i=1}^{k} \mu_{i+1} + \left(S_k + \left(\frac{1}{2} \sum_{i=1}^{k} \mu_{i+1} \right)^2 \right)^{1/2} \leq \sum_{i=1}^{k} \mu_{i+1} + (S_k)^{1/2} \leq \sum_{i=1}^{k+1} \mu_i + C_1 + (F_k)^{1/2} \tag{80}$$

Dropping the term $\|u_K\|^2$ in the inequality (74) and using the bound on $\|u_k\|$ we have:

$$\begin{aligned} &2\alpha_K^{succ} t_K^2 v_K \\ &\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \sum_{k=1}^{K} 2\alpha_k t_k \|(\nabla f(y_k) - g_k)\| \cdot \|u_{k-1}\| \\ &+ \sum_{k=1}^{K} 2\alpha_k t_k^2 |(f_k^y - f(y_k)) + (f(x_k) - f_k^p)| \\ &\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \sum_{k=1}^{K} \mu_k \cdot \|u_{k-1}\| + F_K \\ &\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k} \mu_i + C_1 + (F_k)^{1/2} \right] + F_K \end{aligned}$$

$$
\begin{aligned}
&\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k} \mu_i\right] + C_1 \sum_{k=1}^{K} \mu_k + (F_K)^{1/2} \sum_{k=1}^{K} \mu_k + F_K \\
&\leq 2\alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2 + \sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + \sum_{k=1}^{K} \mu_k^2 + \frac{1}{2} C_1^2 + \left(\sum_{k=1}^{K} \mu_k\right)^2 + \frac{1}{2} F_K + F_K \\
&\leq 3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + 2\sum_{k=1}^{K} \mu_k^2 + \frac{3}{2} F_K \qquad (81)
\end{aligned}
$$

where the last line follows from $\left(\sum_{k=1}^{K} \mu_k\right)^2 = 2\sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + \sum_{k=1}^{K} \mu_k^2$ and the fact that $\frac{1}{2}C_1^2 = \alpha_0^{succ} t_0^2 v_0 + \|u_0\|^2$.

Recall that N_ϵ be the stopping time that $N_\epsilon = \inf\{k : v_k \leq \epsilon\}$. Hence for every $K \leq N_\epsilon - 1$, we have that $v_K > \epsilon$ and

$$
\begin{aligned}
&2\alpha_K^{succ} t_K^2 \\
&\leq \frac{3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + 2\sum_{k=1}^{K} \mu_k^2 + \frac{3}{2} F_K}{v_K} \\
&\leq \frac{3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + 2\sum_{k=1}^{K} \mu_k^2 + \frac{3}{2} F_K}{\epsilon} \qquad (82)
\end{aligned}
$$

Let $\mathcal{N}_{TS}^K$ be the set of true successful iterations with $\alpha_k \geq \bar{\alpha}, k \leq K$ and N_{TS}^K be the cardinality of that set. Using the inequality $\alpha_K^{succ} t_K^2 \geq \left(\sum_{k \in S, k \leq K} \sqrt{\alpha_k}/2\right)^2$ for $K \leq N_\epsilon - 1$ from Lemma 14 we have

$$
\alpha_K^{succ} t_K^2 \geq \left(\sum_{k \in S, k \leq K} \sqrt{\alpha_k}/2\right)^2 \geq \left(\sum_{k \in \mathcal{N}_{TS}^K, k \leq K} \sqrt{\alpha_k}/2\right)^2 \geq \left(N_{TS}^K \sqrt{\bar{\alpha}}/2\right)^2 = \frac{1}{4}(N_{TS}^K)^2 \bar{\alpha}. \qquad (83)
$$

Substituting this to the previous inequality we get the following for $K \leq N_\epsilon - 1$

$$
(N_{TS}^K)^2 \leq \frac{2}{\bar{\alpha}\epsilon} \left[3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\sum_{k=1}^{K} \mu_k \cdot \left[\sum_{i=1}^{k-1} \mu_i\right] + 2\sum_{k=1}^{K} \mu_k^2 + \frac{3}{2} F_K\right] \qquad (84)
$$

Let I_ϵ^K be the event $K \leq N_\epsilon - 1$. Taking total expectation we have:

$$
\begin{aligned}
&\mathbb{E}[(N_{TS}^K)^2] \\
&\leq \frac{2}{\bar{\alpha}\epsilon} \left[3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\mathbb{E}\left[\sum_{k=1}^{K} \mu_k \cdot \left(\sum_{i=1}^{k-1} \mu_i\right)\right] + 2\mathbb{E}\left[\sum_{k=1}^{K} \mu_k^2\right] + \frac{3}{2}\mathbb{E}[F_K]\right] \\
&\leq \frac{2}{\bar{\alpha}\epsilon} \left[3\alpha_0^{succ} t_0^2 v_0 + 2\|u_0\|^2 + 3\mathbb{E}\left(\sum_{k=1}^{K} \sum_{i=1}^{k-1} \mathbb{E}\left[\mu_k \mu_i\right]\right) + 2\mathbb{E}\left(\sum_{k=1}^{K} \mathbb{E}\left[\mu_k^2\right]\right)\right]
\end{aligned}
$$

$$+ \frac{2}{\bar{\alpha}\epsilon}\left[\frac{3}{2}\mathbb{E}\left(\sum_{k=1}^{K}\mathbb{E}\left(2\alpha_k t_k^2 |f_k^y - f(y_k) + f(x_k) - f_k^p|\right)\right)\right] \tag{85}$$

where the last line follows from the general Wald's identity as it satisfies the following conditions:

Lemma 15 (Wald's identity) *Assume that*

- $\{X_n\}_{n\geq 0}$ *are finite-mean random variables.*
- N *is a stopping time for the sequence* $\{X_n\}_{n\geq 0}$
- *The infinite series satisfies*

$$\sum_{n=1}^{\infty}\mathbb{E}[|X_n|1_{\{N\geq n\}}] < \infty.$$

Then we have

$$\mathbb{E}\left[\sum_{n=1}^{N}X_n\right] = \mathbb{E}\left[\sum_{n=1}^{N}\mathbb{E}[X_n]\right]$$

From Assumption 2(i) we have:

$$\mathbb{E}\left[\mu_k^2|\mathcal{F}_{k-1}^{G\cdot F}\right] = \mathbb{E}\left[\alpha_k^2 t_k^2 \|(\nabla f(y_k) - g_k)\|^2|\mathcal{F}_{k-1}^{G\cdot F}\right] \leq \frac{1}{k^{2+\beta}}, \tag{86}$$

which leads to

$$\left(\mathbb{E}\left[\mu_k|\mathcal{F}_{k-1}^{G\cdot F}\right]\right)^2 \leq \mathbb{E}\left[\mu_k^2|\mathcal{F}_{k-1}^{G\cdot F}\right] \leq \frac{1}{k^{2+\beta}}, \tag{87}$$

thus

$$\mathbb{E}\left[\mu_k|\mathcal{F}_{k-1}^{G\cdot F}\right] \leq \frac{1}{k^{1+0.5\beta}}, \tag{88}$$

Taking the total expectation we have

$$\mathbb{E}\left[\mu_k^2\right] \leq \frac{1}{k^{2+\beta}} \text{ and } \mathbb{E}\left[\mu_k\right] \leq \frac{1}{k^{1+0.5\beta}}, \tag{89}$$

Now we consider the quantity $\mathbb{E}\left[\mu_k\mu_i|\mathcal{F}_{k-1}^{G\cdot F}\right]$ where $i < k$. From the definition of $\mathcal{F}_{k-1}^{G\cdot F}$ we have that it contains the information of g_{k-1}, f_{k-1}^y, f_{k-1}^p and y_{k-1}, thus determines the quantities in step 6-10 of Algorithm 1 and includes $\mu_{k-1} = 2\alpha_{k-1}t_{k-1}\|(\nabla f(y_{k-1}) - g_{k-1})\|$.

Hence for every $i < k$ we have $\mu_i \in \mathcal{F}_{k-1}^{G\cdot F}$ and

$$\mathbb{E}\left[\mu_k\mu_i|\mathcal{F}_{k-1}^{G\cdot F}\right] = \mathbb{E}\left[\mu_k|\mathcal{F}_{k-1}^{G\cdot F}\right]\mu_i \tag{90}$$

Now we consider the following term for $i < k \leq K$:

$$\mathbb{E}\left[\mu_k\mu_i\right] = \mathbb{E}\left[\mathbb{E}\left[\mu_k\mu_i|\mathcal{F}_{k-1}^{G\cdot F}\right]\right] \tag{91}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[\mu_k|\mathcal{F}_{k-1}^{G\cdot F}\right]\mu_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{k^{1+0.5\beta}}\mu_i\right]\right] \\
&= \frac{1}{k^{1+0.5\beta}}\mathbb{E}\left[\mathbb{E}\left[\mu_i|\mathcal{F}_{i-1}^{G\cdot F}\right]\right] \\
&= \frac{1}{k^{1+0.5\beta}}\frac{1}{i^{1+0.5\beta}}
\end{aligned}
$$

Using similar derivations in the previous bound (85) we have:

$$
\begin{aligned}
&\mathbb{E}[(N_{TS}^K)^2] \\
&\le \frac{2}{\bar{\alpha}\epsilon}\left[3\alpha_0^{succ}t_0^2 v_0 + 2\left\|u_0\right\|^2 + 3\mathbb{E}\left(\sum_{k=1}^{K}\sum_{i=1}^{k-1}\mathbb{E}\left[\mu_k\mu_i\right]\right) + 2\mathbb{E}\left(\sum_{k=1}^{K}\mathbb{E}\left[\mu_k^2\right]\right)\right] \\
&+ \frac{2}{\bar{\alpha}\epsilon}\left[3\mathbb{E}\left(\sum_{k=1}^{K}\mathbb{E}\left(\alpha_k t_k^2|f_k^y - f(y_k) + f(x_k) - f_k^p|\right)\right)\right] \\
&\le \frac{2}{\bar{\alpha}\epsilon}\left[3\alpha_0^{succ}t_0^2 v_0 + 2\left\|u_0\right\|^2 + 3\mathbb{E}\left(\sum_{k=1}^{K}\frac{1}{k^{1+0.5\beta}}\sum_{i=1}^{k-1}\frac{1}{i^{1+0.5\beta}}\right) + 2\mathbb{E}\left(\sum_{k=1}^{K}\frac{1}{k^{2+\beta}}\right)\right] \\
&+ \frac{2}{\bar{\alpha}\epsilon}\left[\mathbb{E}\left(\sum_{k=1}^{K}\frac{6}{k^{1+\beta'}}\right)\right]. \qquad (92)
\end{aligned}
$$

Note that

$$
\begin{aligned}
\sum_{k=1}^{K}\frac{1}{k^{1+0.5\beta}} < 1 + \sum_{k=2}^{\infty}\frac{1}{k^{1+0.5\beta}} < 1 + \int_1^{\infty}\frac{1}{x^{1+0.5\beta}}dx = \frac{\beta+2}{\beta} \\
\sum_{k=1}^{K}\frac{1}{k^{1+\beta'}} < 1 + \sum_{k=2}^{\infty}\frac{1}{k^{1+\beta'}} < 1 + \int_1^{\infty}\frac{1}{x^{1+\beta'}}dx = \frac{\beta'+1}{\beta'}
\end{aligned}
$$

and

$$
\sum_{k=1}^{K}\frac{1}{k^{2+\beta}} < 1 + \sum_{k=2}^{\infty}\frac{1}{k^{2+\beta}} < 1 + \int_1^{\infty}\frac{1}{x^{2+\beta}}dx = \frac{\beta+2}{\beta+1}
$$

we have the final bound for $\mathbb{E}[(N_{TS}^K)^2]$:

$$
\mathbb{E}[(N_{TS}^K)^2] \le \frac{2}{\bar{\alpha}\epsilon}\left[3\alpha_0^{succ}t_0^2 v_0 + 2\left\|u_0\right\|^2 + \frac{3(\beta+2)^2}{\beta^2} + \frac{2(\beta+2)}{\beta+1} + \frac{6(\beta'+1)}{\beta'}\right]. \qquad (93)
$$

Finally, by the expectation inequality we have:

$$
\mathbb{E}[N_{TS}^K] \le \sqrt{\frac{2}{\bar{\alpha}\epsilon}\left[3\alpha_0^{succ}t_0^2 v_0 + 2\left\|u_0\right\|^2 + \frac{3(\beta+2)^2}{\beta^2} + \frac{2(\beta+2)}{\beta+1} + \frac{6(\beta'+1)}{\beta'}\right]}. \qquad (94)
$$

for every $K \le N_\epsilon - 1$. Thus

$$\mathbb{E}[N_{TS}] \leq \sqrt{\frac{2}{\bar{\alpha}\epsilon}\left[3\alpha_0^{succ}t_0^2v_0 + 2\left\|u_0\right\|^2 + \frac{3(\beta+2)^2}{\beta^2} + \frac{2(\beta+2)}{\beta+1} + \frac{6(\beta'+1)}{\beta'}\right]}. \tag{95}$$

■