Supplemental Materials

A Solving (1) in the main paper directly by SGD is biased

Applying the standard chain rule, we obtain the gradient of the cost function in (1) as

$$\frac{1}{n_X}\sum_{i=0}^{n_X-1}\phi_i'\big(\overline{f}_i(\theta)\big)\overline{f}_i'(\theta) \tag{1}$$

where:

$$\begin{aligned}\overline{f}_i(\theta) &\triangleq \frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_\theta(x_i, y_{ij})\\ \overline{f}_i'(\theta) &\triangleq \frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_\theta'(x_i, y_{ij})\end{aligned} \tag{2}$$

and $f_\theta'\big(x,y\big)$ denotes the $d \times \ell$ matrix, with its $(i,j)^{\text{th}}$ element defined to be:

$$\big[f_\theta'\big(x,y\big)\big]_{i,j} = \frac{\partial}{\partial \theta_i}\big[\big(f_\theta\big(x,y\big)\big]_j \tag{3}$$

Note from (1) that there are empirical averages inside and outside $\phi'(\cdot)$. Therefore, if we sample these empirical averages simultaneously, the stochastic gradient estimator would be biased. In other words, a direct application of stochastic gradient descent to (1) would be intrinsically biased.

B SVRPDA-II algorithm

Algorithm 1 in this supplementary material summarizes the full details of the SVRPDA-II algorithm, which was developed in Section 3.3 of the main paper. Note that it no longer requires the computation or the storage of $\overline{f}_i(\tilde{\theta})$ in (4). Also note that the $\overline{f}_{i_k}(\tilde{\theta})$ in (7) is replaced with $f_{\tilde{\theta}}'(x_{i_k}, y_{i_k j_k''})$ now.

C Experiment details

C.1 Implementation details in risk-averse learning

As we discussed in Section 2, we adopt the alternative formulation (6) for the second term so that it becomes a special case of our general problem (1). Then, using the argument in Section 3.1, the second term in (4) can be rewritten into the objective in (8). Combining it with the first term in (4), the original problem (4) can be reformulated into the following equivalent min-max form:

$$\min_{\theta\in\mathbb{R}^d}\max_{w}\frac{1}{n}\sum_{i=0}^{n-1}\bigg(\Big\langle\frac{1}{n}\sum_{j=0}^{n-1}\langle x_i - x_j, \theta\rangle, w_i\Big\rangle - \phi^*(w_i) - \langle x_i, \theta\rangle\bigg) \tag{10}$$

where $w_i \in \mathbb{R}$, $\phi^*(w_i) = w_i^2/4$ and $w = \{w_0, \ldots, w_{n-1}\}$. Note that the above min-max problem has an extra $\langle x_i, \theta\rangle$ term within the sum. Since it is in a standard empirical average form, we can deal with it in a straightforward manner. Notice that (10) is exactly of the form (1) in the main paper except the last term $\langle x_i, \theta\rangle$ within the summation, which as we will show next, can be dealt with in a straightforward manner.

Taking out the $\langle x_i, \theta\rangle$ term in (10), based on the discussion in Section 3 of the main paper, the batch gradients used in the algorithm are as follows. Batch gradient of (10) with respect to w_i, for each $0 \le i \le n-1$ can be written as:

$$\overline{f}_i(\theta) = \frac{1}{n}\sum_{j=0}^{n-1} f_\theta(x_i, x_j) = \frac{1}{n}\sum_{j=0}^{n-1}\langle x_i - x_j, \theta\rangle \tag{11}$$

Algorithm 1 SVRPDA-II

```
Inputs: data $\{(x_i, y_{ij}) : 0 \le i < n_X, 0 \le j < n_{Y_i}\}$; step-sizes $\alpha_\theta$ and $\alpha_w$; # inner iterations $M$.
Initialization: $\tilde{\theta}_0 \in \mathbb{R}^d$ and $\tilde{w}_0 \in \mathbb{R}^{\ell n_X}$.
for $s = 1, 2, \dots$ do
    Set $\tilde{\theta} = \tilde{\theta}_{s-1}$, $\theta^{(0)} = \tilde{\theta}$, $w^{(0)} = \tilde{w}_{s-1}$, and compute the batch quantities (for each $0 \le i < n_X$):
        $$U_0 = \sum_{i=0}^{n_X-1} \sum_{j=0}^{n_{Y_i}-1} \frac{f'_{\tilde{\theta}}(x_i, y_{ij}) w_i^{(0)}}{n_X n_{Y_i}}, \quad \overline{f}_i(\tilde{\theta}) \triangleq \sum_{j=0}^{n_{Y_i}-1} \frac{f_{\tilde{\theta}}(x_i, y_{ij})}{n_{Y_i}}. \qquad (4)$$
    for $k = 1$ to $M$ do
        Randomly sample $i_k \in \{0, \dots, n_X - 1\}$ and then $j_k \in \{0, \dots, n_{Y_{i_k}} - 1\}$ at uniform.
        Compute the stochastic variance reduced gradient for dual update:
            $$\delta_k^w = f_{\theta^{(k-1)}}(x_{i_k}, y_{i_k j_k}) - f_{\tilde{\theta}}(x_{i_k}, y_{i_k j_k}) + \overline{f}_{i_k}(\tilde{\theta}). \qquad (5)$$
        Update the dual variables:
            $$w_i^{(k)} = \begin{cases} \arg\min_{w_i} \Big[ -\langle \delta_k^w, w_i \rangle + \phi_i^*(w_i) + \frac{1}{2\alpha_w} \|w_i - w_i^{(k-1)}\|^2 \Big] & \text{if } i = i_k \\ w_i^{(k-1)} & \text{if } i \ne i_k \end{cases}. \qquad (6)$$
        Update $U_k$ according to the following recursion:
            $$U_k = U_{k-1} + \frac{1}{n_X} f'_{\tilde{\theta}}(x_{i_k}, y_{i_k j''_k}) \big( w_{i_k}^{(k)} - w_{i_k}^{(k-1)} \big). \qquad (7)$$
        Randomly sample $i'_k \in \{0, \dots, n_X - 1\}$ and then $j'_k \in \{0, \dots, n_{Y_{i'_k}} - 1\}$, independent of $i_k$ and $j_k$, and compute the stochastic variance reduced gradient for primal update:
            $$\delta_k^\theta = f'_{\theta^{(k-1)}}(x_{i'_k}, y_{i'_k j'_k}) w_{i'_k}^{(k)} - f'_{\tilde{\theta}}(x_{i'_k}, y_{i'_k j'_k}) w_{i'_k}^{(k)} + U_k. \qquad (8)$$
        Update the primal variable:
            $$\theta^{(k)} = \arg\min_{\theta} \Big[ \langle \delta_k^\theta, \theta \rangle + g(\theta) + \frac{1}{2\alpha_\theta} \|\theta - \theta^{(k-1)}\|^2 \Big]. \qquad (9)$$
    end for
    Option I: Set $\tilde{w}_s = w^{(k)}$ and $\tilde{\theta}_s = \theta^{(k)}$.
    Option II: Set $\tilde{w}_s = w^{(k)}$ and $\tilde{\theta}_s = \theta^{(t)}$ for randomly sampled $t \in \{0, \dots, M-1\}$.
end for
Output: $\tilde{\theta}_s$ at the last outer-loop iteration.
```

Batch gradient of (10) with respect to θ (without the $\langle x_i, \theta \rangle$ term) is given by:

$$\begin{aligned} L'_\theta(\theta, w) &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} f'_\theta\big(x_i, x_j\big) w_i \\ &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \big(x_i - x_j\big) w_i \end{aligned} \tag{12}$$

For each $0 \le i \le n-1$, gradient of $\overline{f}_{\theta,i}(x_i)$ is given by:

$$\overline{f}'_i(\theta) \triangleq \frac{1}{n} \sum_{j=0}^{n-1} f'_\theta(x_i, x_j) = x_i - \frac{1}{n} \sum_{j=0}^{n-1} x_j \tag{13}$$

Based on the above derivation and the expression (17) in the main paper, the stochastic variance reduced gradient for the dual update in both SVRPDA-I and SVRPDA-II is given by

$$\delta_k^w = \langle x_i - x_j, \theta \rangle + \overline{f}_{\tilde{\theta}}(x_i) \tag{14}$$

and the stochastic variance reduced gradient for the primal update is given by

$$\delta_k^\theta \triangleq (x_i - x_j)w_i - (x_i - x_j)w_i + \overline{L}'_\theta(\tilde{\theta}, w) = \overline{L}'_\theta(\tilde{\theta}, w) \tag{15}$$

Note that, since the function f_θ is linear in θ, the variance reduced gradient for the primal variable is in-fact the full batch gradient.

Next, due to the additional $\langle x_i, \theta \rangle$ term in (10) (which was ommitted in the above definitions), there is an additional term that needs to be added to the variance reduced gradient in (15). Denoting $g_\theta(x_i) = \langle x_i \, , \theta \rangle$ and $g'_\theta(x_i) = x_i$ the correction batch term is given by:

$$\overline{g}'_\theta = \frac{1}{n} \sum_{i=0}^{n-1} g'_\theta(x_i) = \frac{1}{n} \sum_{i=0}^{n-1} x_i \tag{16}$$

which is independent of θ. In summary, the final variance reduced stochastic gradient for the primal update in both SVRPDA-I and SVRPDA-II is given by:

$$\delta_k^\theta = \overline{L}'_\theta(\tilde{\theta}, w) - \frac{1}{n} \sum_{i=0}^{n-1} x_i \tag{17}$$

C.2 Hyper-parameter choices for algorithms

In this subsection, we provide the hyper-parameters that are used in our experiments on risk-averse learning (Section 5). We first list the hyper-parameters of our methods below:

- SVRPDA-I: $M = n$, $\alpha_\theta = 0.0003$, $\alpha_w = 100$.
- SVRPDA-II: $M = n$, $\alpha_\theta = 0.0003$, $\alpha_w = 100$.

Then, we provide the hyper-parameters used in the baseline methods:

- Compositional-SVRG-1 (Algorithm 2 of [1]): $K = n$, $A = 6$, $\gamma = 0.0003$;
- Compositional-SVRG-2 (Algorithm 3 of [1]): $K = n$, $A = 3$, $B = 3$, $\gamma = 0.0004$;
- ASCVRG: The results are obtained by using their publicly released code on github: `https://github.com/tyDLin/SCVRG` with the same setting and choice of hyper-parameters.
- Batch gradient descent: step size $\alpha = 0.01$;

Note that, for the Compotional-SVRGs and batch gradient algorithms, the above choice of the hyper-parameters are obtained by sweeping through a set of hyper-parameters and choosing the ones with the best performance. For ASCVRG, we use the the publicly released code by the authors.

D Additional experiments on MDP policy evaluation

Consider a Markov decision process (MDP) problem with state space S and action space A. We assume that both S and A are finite, and define $\mathsf{S} = \{1, \ldots, S\}$. For any $1 \le i, j \le S$, we denote by $r_{i,j}$ the reward associated with transition from state i to state j. Given a policy $\pi : \mathsf{S} \to \mathcal{P}(\mathsf{A})$, where $\mathcal{P}(\mathsf{A})$ denote the probability space over A, we let $P^\pi \in \mathbb{R}^{S \times S}$ denote the associated state transition probability matrix. The goal in policy evaluation is to estimate the value function $V^\pi : \mathsf{S} \to \mathbb{R}$ associated with the policy π, which is a fixed-point solution to the following Bellman equation:

$$V^\pi(i) = \sum_{j=1}^{S} P^\pi_{i,j} \big(r_{i,j} + \gamma V^\pi(j) \big) \,, \qquad 1 \le i \le S,$$

where $0 < \gamma < 1$ denotes the discount factor. We consider a linear function approximation to the value function: $V^\pi(i) \approx \langle \Psi_i \, , \theta \rangle$, where $\{\Psi_i \in \mathbb{R}^d : 1 \le i \le S\}$ denotes the feature vectors, and $\theta \in \mathbb{R}^d$ denotes the weight vector to be learned. The problem of finding the optimal weight vector θ^* that best approximates V^π can be formulated as the following optimization problem [2, 4]:

$$\theta^* = \arg\min_\theta \left\{ F(\theta) := \frac{1}{S} \sum_{i=1}^{S} \Big(\langle \Psi_i \, , \theta \rangle - \sum_{j=1}^{S} P^\pi_{i,j} \big(r_{i,j} + \gamma \langle \Psi_j \, , \theta \rangle \big) \Big)^2 \right\}. \tag{18}$$

Note that the above problem can be expressed as a special case of (1) by using the following identifications: For each $1 \le i \le S$, $n_X = n_{Y_i} = S$, $\phi_i(u) = u^2$, and

$$f_\theta(x_i, y_{ij}) = \langle \Psi_i\,, \theta\rangle - S \cdot P^\pi_{i,j}\big(r_{i,j} + \gamma\langle \Psi_j\,, \theta\rangle\big)\,, \qquad 1 \le i, j \le S. \tag{19}$$

And it is also possible, although less intuitive, to rewrite (18) as a special case of (2). In existing composition optimization literature such as [4, 2], this is achieved via higher dimensional transformation, with $f_\theta : \mathbb{R}^d \to \mathbb{R}^{2S}$, $\phi_i : \mathbb{R}^{2S} \to \mathbb{R}$, $n_Y = S$, and $n_X = S$. Denote by Q^π_θ the Q-function:

$$Q^\pi_\theta(i) := \sum_{j=1}^{S} P^\pi_{i,j}\big(r_{i,j} + \gamma\langle \Psi_j\,, \theta\rangle\big)\,, \qquad 1 \le i \le S.$$

Then, by defining the function f_θ and ϕ_i such that

$$\frac{1}{S}\sum_{j=1}^{S} f_\theta(y_j) = \Big[\langle \Psi_1, \theta\rangle\,, Q^\pi_\theta(1)\,, \dots\,, \langle \Psi_S, \theta\rangle\,, Q^\pi_\theta(S)\Big] \tag{20}$$

$$\frac{1}{S}\sum_{i=1}^{S} \phi_i\left(\Big[\langle \Psi_1, \theta\rangle\,, Q^\pi_\theta(1)\,, \dots\,, \langle \Psi_S, \theta\rangle\,, Q^\pi_\theta(S)\Big]\right) = \frac{1}{S}\sum_{i=1}^{S}\Big(\langle \Psi_i, \theta\rangle - Q^\pi_\theta(i)\Big)^2,$$

the problem (18) can be reformulated as (2). The reader is referred to [4, 2] for more details.

We evaluate our algorithms on two experimental settings, one with $S = 10$, $d = 5$, and another with $S = 10^4$ and $d = 10$. In each of these two cases, γ was set to be 0.9, and both the transition probability matrix $P^\pi \in \mathbb{R}^{S\times S}$ and the feature vectors $\{\Psi_i \in \mathbb{R}^d : 1 \le i \le S\}$ were randomly generated. We compare the performance our algorithms SVRPDA-I and SVRPDA-II with the C-SAGA algorithm of [4], as it is the most recent composition optimization that we are aware of, and is shown to be superior to all existing algorithms on this MDP policy evaluation task [4]. The hyper-parameters for the C-SAGA algorithm were chosen as follows. For $S = 10$ case, we choose all hyper-parameters as in [4]: Mini-batch size $s = 1$ and step-size $\eta = 0.1$. For the $S = 10^4$ case, we choose mini-batch size $s = 100$, and step-size $\eta = 0.0005$ (see [4] for details on what these hyper-parameters mean). The hyper-parameters for the SVRPDA-I and SVRPDA-II are chosen as follows. For the MDP with $S = 10$, we choose $M = 150$, $\alpha_\theta = 0.1$, and $\alpha_w = 0.25$ for SVRPDA-I, and choose $M = 15$, $\alpha_\theta = 0.5$, and $\alpha_w = 1.25$ for SVRPDA-II. For the MDP with $S = 10^4$, we choose $M = 13500$, $\alpha_\theta = 0.01$ and $\alpha_w = 16 \times 10^4$ for SVRPDA-I. For SVRPDA-II, $M = 1350$, $\alpha_\theta = 0.01$ and $\alpha_w = 16 \times 10^4$.

The performance criteria was chosen to be the number of oracle calls required to achieve a certain "objective gap", defined as $F(\theta) - F(\theta^*)$. Notice that "one function call", when the function is of the form (19) *is not comparable* to one function call, when the function is defined according to (20). Due to the fact that f_θ for the C-SAGA algorithm is of dimension $2S$ (as opposed to 1 in our formulation), we count $2S$ oracle calls whenever a function of this form is called for the purpose of fair comparison. The results for MDP with $S = 10$ and $S = 10^4$ are reported in Figures 1 and 2, respectively. We observe that despite having much smaller memory requirement, SVRPDA-II has a comparable/better performance than C-SAGA, while SVRPDA-I is clearly better than C-SAGA.

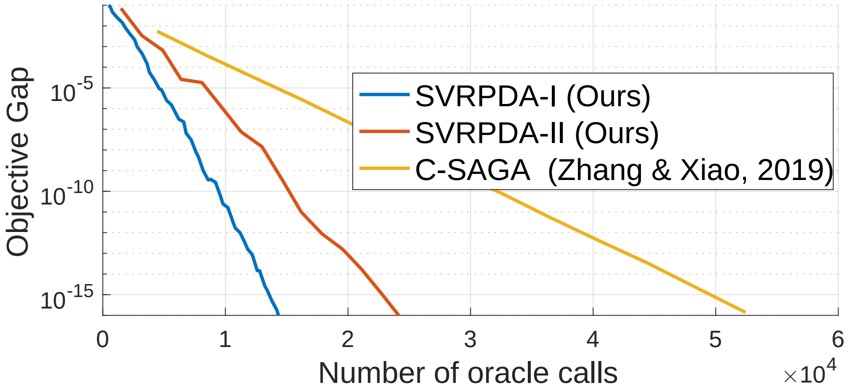

Figure 1: Performance comparison of our algorithms with the C-SAGA on the MDP with $S = 10$. The performance is measured in terms of the number of oracle calls to achieve a certain objective gap.

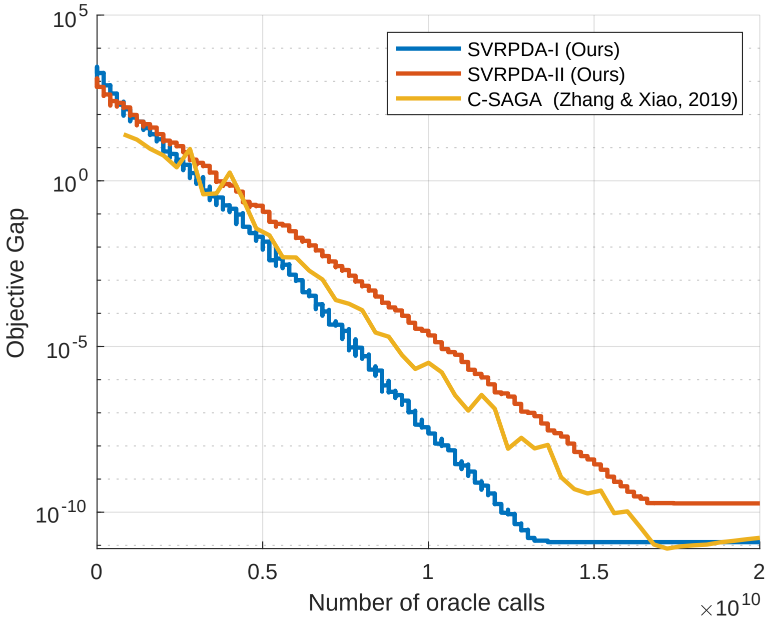

Figure 2: Performance comparison of our algorithms with the C-SAGA on the MDP with $S = 10^4$. The performance is measured in terms of the number of oracle calls to achieve a certain objective gap.

E Convergence and complexity of SVRPDA-I: Proof

In this section, we derive the (non-asymptotic) bound of the SVRPDA-I algorithm and its total computation complexity. For convenience, we first repeat the saddle point formulation and the definition of several quantities below:

$$\min_{\theta} \max_{w} \frac{1}{n_X} \sum_{i=0}^{n_X-1} \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Big(\langle f_\theta(x_i, y_{ij}),\, w_i \rangle - \phi_i^*(w_i) \Big) + g(\theta) \tag{21}$$

Also, recall the definitions of $\overline{f}_i(\theta)$ and $\overline{f}'_i(\theta)$:

$$\overline{f}_i(\theta) = \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} f_\theta(x_i, y_{ij}), \quad \overline{f}'_i(\theta) = \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} f'_\theta(x_i, y_{ij}) \tag{22}$$

Furthermore, we defined $L(\theta, w)$ and its gradient as

$$L(\theta, w) := \frac{1}{n_X} \sum_{i=0}^{n_X-1} \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Big(\langle f_\theta(x_i, y_{ij}),\, w_i \rangle - \phi_i^*(w_i) \Big) \tag{23}$$

$$= \frac{1}{n_X} \sum_{i=0}^{n_X-1} \Big(\langle \overline{f}_i(\theta),\, w_i \rangle - \phi_i^*(w_i) \Big) \tag{24}$$

$$L'_\theta(\theta, w) := \frac{1}{n_X} \sum_{i=0}^{n_X-1} \overline{f}'_i(\theta) w_i \tag{25}$$

Using the above notations, the saddle point problem (21) can be rewritten as

$$\min_{\theta} \min_{w} \frac{1}{n_X} \sum_{i=0}^{n_X-1} \Big(\langle \overline{f}_i(\theta), w_i \rangle - \phi_i^*(w_i) \Big) + g(\theta) \tag{26}$$

E.1 Compact Notation

Throughout this section, we introduce the following compact notation, to ease exposition of the proof: For any $\theta \in \mathbb{R}^d$, and $0 \le i \le n_X - 1$ and $0 \le j \le n_{Y_i} - 1$, we denote:

$$f_{ij}(\theta) \equiv f_\theta(x_i, y_{ij}) \tag{27}$$

Therefore, the stochastic variance reduced gradient for dual update defined in (12) of the main paper is rewritten as:

$$\delta_k^w = f_{i_k j_k}(\theta^{(k-1)}) - f_{i_k j_k}(\widetilde{\theta}) + \overline{f}_{i_k}(\widetilde{\theta}) \tag{28}$$

Similarly, the stochastic variance reduced gradient for primal update defined in (15) of the main paper is:

$$\delta_k^\theta = f'_{i'_k j'_k}(\theta^{(k-1)}) w_{i'_k}^{(k)} - f'_{i'_k j'_k}(\widetilde{\theta}) w_{i'_k}^{(k)} + L'_\theta(\widetilde{\theta}, w^{(k)}) \tag{29}$$

where we used the fact that $U_k \equiv L'_\theta(\widetilde{\theta}, w^{(k)})$. We now proceed to recall the Algorithm 1 rewritten in a simplified form, using the compact notation.

E.2 Algorithm

Before we proceed to prove the convergence of the algorithm, we first recall the update equations of the algorithm. The following updates are at stage s of the outerloop; To simplify exposition, we suppress dependency on s, and let $\widetilde{\theta} \equiv \widetilde{\theta}_s$ throughout.

For the dual update, at each iteration k, we first randomly pick an index $0 \le i_k \le n_X - 1$ at uniform, and then pick another index $0 \le j_k \le n_{Y_{i_k}} - 1$ at uniform. For the chosen (i_k, j_k), we first compute the variance reduced stochastic gradient δ_k^w of $\overline{f}_i(\theta)$ using (28):

$$\delta_k^w = f_{i_k j_k}(\theta^{(k-1)}) - f_{i_k j_k}(\widetilde{\theta}) + \overline{f}_{i_k}(\widetilde{\theta})$$

Then, we update the dual variables according to the recursion (13):

$$w_i^{(k)} = \begin{cases} \arg\min\limits_{w_i} \Big[-\Big\langle \delta_k^w,\, w_i - w_i^{(k-1)} \Big\rangle + \phi_i^*(w_i) + \dfrac{1}{2\alpha_w} \|w_i - w_i^{(k-1)}\|^2 \Big] & \text{if } i = i_k \\ w_i^{(k-1)} & \text{if } i \neq i_k \end{cases} \tag{30}$$

For the primal update, at iteration k, we randomly pick another independent set of indices (i'_k, j'_k) with $0 \le i'_k \le n_X - 1$ and $0 \le j'_k \le n_{Y_{i_k}} - 1$, and compute the variance reduced stochastic gradient δ_k^θ of $L(\theta, w)$ with respect to θ using (29):

$$\delta_k^\theta = f'_{i'_k j'_k}(\theta^{(k-1)}) w_{i'_k}^{(k)} - f'_{i'_k j'_k}(\widetilde{\theta}) w_{i'_k}^{(k)} + L'_\theta(\widetilde{\theta}, w^{(k)})$$

Then, we update the primal variable θ according to the recursion (16):

$$\theta^{(k)} = \arg\min_\theta \Big\{ \langle \delta_k^\theta, \theta \rangle + g(\theta) + \frac{1}{2\alpha_\theta} \|\theta - \theta^{(k-1)}\|^2 \Big\} \tag{31}$$

E.3 Assumptions

We restate the Assumptions in Section 4 here using the notation in (27) to make the reading easier:

Assumption E.1. *The function $g(\theta)$ is μ-strongly convex in θ, and each ϕ_i is $1/\gamma$-smooth.*

Assumption E.2. *The merit functions $\phi_i(u)$ are Lipschitz with a uniform constant B_w:*

$$|\phi_i(u) - \phi_i(u')| \le B_w \|u - u'\|, \quad \forall u, u' \in \mathbb{R}^\ell, 0 \le i \le n_X - 1.$$

Assumption E.3. *$f_{ij}(\theta)$ is B_θ-smooth in θ, and has bounded gradients with constant B_f: For each $0 \le i \le n_X - 1$ and $0 \le j \le n_{Y_i} - 1$,*

$$\|f'_{ij}(\theta_1) - f'_{ij}(\theta_2)\| \le B_\theta \|\theta_1 - \theta_2\|, \quad \|f'_{ij}(\theta)\| \le B_f, \quad \forall \theta, \theta_1, \theta_2 \in \mathbb{R}^d$$

Assumption E.4. *For each given w in its domain, the function $L(\theta, w)$ defined in* (8) *is convex in θ:*

$$L(\theta_1, w) - L(\theta_2, w) \ge \langle L'_\theta(\theta_2, w),\ \theta_1 - \theta_2 \rangle.$$

E.4 Preliminary results

In this subsection, we introduce lemmas which lay the foundation for the proof of the main convergence result that follows. First, our proof relies on the following important lemma, which is a slightly adjusted version of Lemma 3 in [3] for our problem setting.

Lemma E.5. *Consider any function of the form $P(x) = f(x) + g(x)$, with $x \in \mathbb{R}^d$. Suppose $f(x)$ is linear in x, and $g(x)$ is μ_g-strongly convex. Then, for $\alpha > 0$, the following holds for any vector $v \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$:*

$$P(y) \geq P(x^{(+)}) + \frac{1}{\alpha}\langle (x - x^{(+)}), (y - x)\rangle + \langle (v - f'(x)), (x^{(+)} - y)\rangle + \frac{1}{\alpha}\|x - x^{(+)}\|^2 + \frac{\mu_g}{2}\|y - x^{(+)}\|^2$$

where:

$$\begin{aligned} x^{(+)} &= \mathrm{prox}_{\alpha g}\{x - \alpha v\} \\ &= \arg\min_{w} \left\{ g(w) + \frac{1}{2\alpha}\|w - x + \alpha v\|^2 \right\} \end{aligned}$$

Proof. Based on the definition of $x^{(+)}$, the optimality condition associated with the proximal operator states that there exists a sub-gradient $\xi \in \partial g(x^{(+)})$ such that

$$\frac{x^{(+)} - x + \alpha v}{\alpha} + \xi = 0 \tag{32}$$

where $\partial g(x^{(+)})$ denotes the sub-differential of g at $x^{(+)}$. Next, by the linearity of f and the strong convexity of g, we have, for any $x, y \in \mathbb{R}^d$,

$$\begin{aligned} P(y) &= f(y) + g(y) \\ &\overset{(a)}{\geq} f(x^{(+)}) + \langle f'(x), (y - x^{(+)})\rangle + g(x^{(+)}) + \langle \xi, (y - x^{(+)})\rangle + \frac{\mu_g}{2}\|y - x^{(+)}\|^2 \\ &\overset{(b)}{=} P(x^{(+)}) + \langle f'(x), (y - x^{(+)})\rangle + \langle \xi, (y - x^{(+)})\rangle + \frac{\mu_g}{2}\|y - x^{(+)}\|^2 \\ &\overset{(c)}{=} P(x^{(+)}) + \langle f'(x), (y - x^{(+)})\rangle - \frac{1}{\alpha}\langle (x^{(+)} - x + \alpha v), (y - x^{(+)})\rangle + \frac{\mu_g}{2}\|y - x^{(+)}\|^2 \\ &\overset{(d)}{=} P(x^{(+)}) + \langle f'(x) - v, (y - x^{(+)})\rangle - \frac{1}{\alpha}\langle (x^{(+)} - x), (y - x^{(+)})\rangle + \frac{\mu_g}{2}\|y - x^{(+)}\|^2 \\ &\overset{(e)}{=} P(x^{(+)}) + \langle f'(x) - v, (y - x^{(+)})\rangle + \frac{1}{\alpha}\|x^{(+)} - x\|^2 \\ &\quad - \frac{1}{\alpha}\langle (x^{(+)} - x), (y - x)\rangle + \frac{\mu_g}{2}\|y - x^{(+)}\|^2 \end{aligned} \tag{33}$$

where step (a) follows from the linearity of f and the strong convexity of the function g, step (b) uses the definition $P(x^{(+)}) = f(x^{(+)}) + g(x^{(+)})$, step (c) substitutes the expression of ξ from (32), step (d) rearrange the second and the third terms, and step (e) completes the proof by adding and subtracting x in the second inner product. ☐

The difference between our Lemma E.5 and Lemma 3 in [3] is that our function $f(x)$ is linear (instead of being strongly convex) in x, which is the setting that we are mainly interested in (i.e., linear dependency on the dual variables). As a result, our α can be any positive number (it is constrained to be smaller than a certain positive number in Lemma 3 of [3]). This lemma is useful for deriving a bound when the update recursions are defined by a proximal mapping with an arbitrary update vector v. This is particularly helpful for our case as both our primal and dual updates are in proximal mapping form with the update vector v being variance reduced stochastic gradient.

Next, we quote the Lemma 2 of [3] below:

Lemma E.6. *Let R be a closed convex function on $\mathbb{R}^d$ and let $x, y \in \mathbf{dom}(R)$. Then:*

$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\| \leq \|x - y\| \tag{34}$$

We next introduce a useful property of the conjugate function:

Lemma E.7. *When Assumption E.2 holds, the domain of $\phi_i^*(w_i)$, denoted as* $\mathrm{domain}(\phi_i)$*, satisfies*

$$\mathrm{domain}(\phi_i) \subseteq \{w_i : \|w_i\| \le B_w\} \tag{35}$$

That is, for any w_i that satisfies $\|w_i\| > B_w$, we will have $\phi_i^*(w_i) = +\infty$. In consequence, the dual variables $w_i^{(k)}$ obtained from the dual update (30) *will always be bounded by B_w throughout the iterations.*

Proof. For any given w_i that satisfies $\|w_i\| > B_w$, define $u_i^t = u_i + \frac{w_i}{\|w_i\|}t$, where t is an arbitrary real scalar. Then, by the definition of conjugate function, we have

$$\begin{aligned}
\phi_i^*(w_i) &= \sup_{u_i} \Big[\langle w_i, u_i\rangle - \phi_i(u_i)\Big] \\
&\overset{(a)}{\ge} \sup_t \Big[\Big\langle w_i, u_i + \frac{w_i}{\|w_i\|}t\Big\rangle - \phi_i\Big(u_i + \frac{w_i}{\|w_i\|}t\Big)\Big] \\
&= \sup_t \Big[\|w_i\|t - \phi_i\Big(u_i + \frac{w_i}{\|w_i\|}t\Big) + \phi_i(u_i)\Big] + \langle w_i, u_i\rangle - \phi_i(u_i) \\
&\overset{(b)}{\ge} \sup_t \Big[\|w_i\|t - B_w\Big\|\frac{w_i}{\|w_i\|}t\Big\|\Big] + \langle w_i, u_i\rangle - \phi_i(u_i) \\
&= \sup_t \Big[(\|w_i\| - B_w)t\Big] + \langle w_i, u_i\rangle - \phi_i(u_i) \\
&= +\infty
\end{aligned} \tag{36}$$

where step (a) uses the fact that the supremum over a subset (line) is smaller, and step (b) uses the following inequality obtained from the Lipschitz property of $\phi_i(u_i)$:

$$|\phi_i(u) - \phi_i(u')| \le B_w\|u - u'\|, \quad \forall u, u' \quad \Rightarrow \quad -\phi_i(u) + \phi_i(u') \ge -B_w\|u - u'\| \tag{37}$$

□

E.5 Dual Bound

In order to derive the bound for the dual update, we first introduce an auxiliary dummy variable w'_{ij}:

$$\begin{aligned}
w'_{ij} &= \arg\min_{w_i} \Big[-\Big\langle \delta_{ij}^w,\, w_i - w_i^{(k-1)}\Big\rangle + \phi_i^*(w_i) + \frac{1}{2\alpha_w}\|w_i - w_i^{(k-1)}\|^2\Big] \\
&= \mathrm{prox}_{\alpha_w\phi_i^*}\Big[w_i^{(k-1)} - \alpha_w\delta_{ij}^w\Big]
\end{aligned} \tag{38}$$

where,

$$\delta_{ij}^w := f_{ij}(\theta^{(k-1)}) - f_{ij}(\widetilde{\theta}) + \overline{f}_i(\widetilde{\theta}) \tag{39}$$

The variable w'_{ij} can be understood as the updated value of the dual variable if i and j is selected.

Our analysis in this section focuses on deriving bounds for the $\|w^{(k)} - w^*\|^2$. We will first examine $\|w'_{ij} - w_i^*\|^2$ and then relate it to $\|w^{(k)} - w^*\|^2$. To begin with, for each i and j, we have

$$\begin{aligned}
\|w'_{ij} - w_i^*\|^2 &= \|w'_{ij} - w_i^{(k-1)} + w_i^{(k-1)} - w_i^*\|^2 \\
&= \|w'_{ij} - w_i^{(k-1)}\|^2 + \|w_i^{(k-1)} - w_i^*\|^2 + 2\langle(w'_{ij} - w_i^{(k-1)}), (w_i^{(k-1)} - w_i^*)\rangle
\end{aligned} \tag{40}$$

Now, we upper bound the first and the third terms in (40) together. For a given $\theta^{(k)}$ and i, define

$$P_{w_i}(x) := -\langle\overline{f}_i(\theta^{(k-1)}), x\rangle + \phi_i^*(x) \tag{41}$$

Note that the first part of the function is linear in x and the second part of the function is γ-strongly convex (since ϕ_i is $1/\gamma$-smooth by Assumption E.1). Furthermore, by (38), the update rule for the dummy variables w'_{ij} is defined by a proximal operator. Therefore, we can apply Lemma E.5 with $P(x) \equiv P_{w_i}(x)$ and the following identifications:

$$f(x) = -\langle\overline{f}_i(\theta^{(k-1)}), x\rangle \quad g(x) = \phi_i^*(x) \quad v = -\delta_{ij}^w \quad x = w_i^{(k-1)} \quad x^{(+)} = w'_{ij} \quad y = w_i^* \quad \alpha = \alpha_w$$

which leads to

$$\begin{aligned}
-\big\langle \overline{f}_i(\theta^{(k-1)}), w_i^* \big\rangle + \phi_i^*(w_i^*) \geq & -\big\langle \overline{f}_i(\theta^{(k-1)}), w'_{ij} \big\rangle + \phi_i^*(w'_{ij}) \\
& + \frac{1}{\alpha_w} \langle (w_i^{(k-1)} - w'_{ij}), (w_i^* - w_i^{(k-1)}) \rangle \\
& - \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\, , w'_{ij} - w_i^* \big\rangle \\
& + \frac{1}{\alpha_w} \| w_i^{(k-1)} - w'_{ij} \|^2 + \frac{\gamma}{2} \| w_i^* - w'_{ij} \|^2
\end{aligned} \tag{42}$$

Furthermore, by definition, since w_i^* is the optimal solution to the following optimization problem,

$$w_i^* = \arg\min_{w_i} \; \Big\{ \phi_i^*(w_i) - \big\langle \overline{f}_i(\theta^*), w_i \big\rangle \Big\}$$

and by the fact that the cost function inside the above $\arg\min$ is γ-strongly convex due to $\phi_i^*(\cdot)$, we have

$$-\Big\langle \overline{f}_i(\theta^*),\, w'_{ij} \Big\rangle + \phi_i^*(w'_{ij}) \geq -\Big\langle \overline{f}_i(\theta^*),\, w_i^* \Big\rangle + \phi_i^*(w_i^*) + \frac{\gamma}{2} \| w'_{ij} - w_i^* \|^2 \tag{43}$$

Adding (42) and (43) cancels the ϕ_i^* terms and leads to

$$\begin{aligned}
\big\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w'_{ij} - w_i^* \big\rangle \geq & \frac{1}{\alpha_w} \langle (w_i^{(k-1)} - w'_{ij}), (w_i^* - w_i^{(k-1)}) \rangle \\
& - \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\, , w'_{ij} - w_i^* \big\rangle \\
& + \frac{1}{\alpha_w} \| w_i^{(k-1)} - w'_{ij} \|^2 + \gamma \| w_i^* - w'_{ij} \|^2
\end{aligned} \tag{44}$$

Multiplying both sides by $2\alpha_w$ and rearranging the terms, we obtain

$$\begin{aligned}
& 2\alpha_w \big\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w'_{ij} - w_i^* \big\rangle + 2\alpha_w \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\, , w'_{ij} - w_i^* \big\rangle - 2\alpha_w \gamma \| w'_{ij} - w_i^* \|^2 \\
& \quad \geq 2 \langle (w_i^{(k-1)} - w'_{ij}), (w_i^* - w_i^{(k-1)})) \rangle\rangle + 2 \| w_i^{(k-1)} - w'_{ij} \|^2
\end{aligned} \tag{45}$$

Now, observe that inequality (45) could be used as an uppper bound for the first and third terms on the right hand side of (40). Using this, (40) becomes:

$$\begin{aligned}
& \| w'_{ij} - w_i^* \|^2 \\
& \quad = \| w_i^{(k-1)} - w_i^* \|^2 + \| w'_{ij} - w_i^{(k-1)} \|^2 + 2 \langle (w'_{ij} - w_i^{(k-1)}), (w_i^{(k-1)} - w_i^*) \rangle \\
& \quad \overset{(a)}{=} \| w_i^{(k-1)} - w_i^* \|^2 - \| w'_{ij} - w_i^{(k-1)} \|^2 + 2 \| w'_{ij} - w_i^{(k-1)} \|^2 + 2 \langle (w'_{ij} - w_i^{(k-1)}), (w_i^{(k-1)} - w_i^*) \rangle \\
& \quad \leq \| w_i^{(k-1)} - w_i^* \|^2 - \| w'_{ij} - w_i^{(k-1)} \|^2 + 2\alpha_w \big\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w'_{ij} - w_i^* \big\rangle \\
& \quad + 2\alpha_w \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\, , w'_{ij} - w_i^* \big\rangle - 2\alpha_w \gamma \| w'_{ij} - w_i^* \|^2
\end{aligned} \tag{46}$$

where step (a) added and subtracted a $\| w_i^{(k-1)} - w'_{ij} \|^2$ in order to apply (45) in the following inequality. Dividing both sides by $2\alpha_w$ and combining common terms, we get the following bound:

$$\begin{aligned}
\Big(\frac{1}{2\alpha_w} + \gamma \Big) \| w'_{ij} - w_i^* \|^2 \leq & \frac{1}{2\alpha_w} \| w_i^{(k-1)} - w_i^* \|^2 - \frac{1}{2\alpha_w} \| w'_{ij} - w_i^{(k-1)} \|^2 \\
& + \big\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w'_{ij} - w_i^* \big\rangle + \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\, , w'_{ij} - w_i^* \big\rangle
\end{aligned} \tag{47}$$

Next, we will bound the last term in (47). Consider the full batch dual ascent algorithm. In this case, for each $0 \leq i \leq n_X - 1$, the update rule is given by:

$$\begin{aligned}
\overline{w}_i^{(k)} &= \arg\min_{w_i} \Big[-\Big\langle \overline{f}_i(\theta^{(k-1)}),\, w_i \Big\rangle + \phi_i^*(w_i) + \frac{1}{2\alpha_w} \| w_i - w_i^{(k-1)} \|^2 \Big] \\
&= \operatorname{prox}_{\alpha_w \phi_i^*} \Big[w_i^{(k-1)} - \alpha_w \overline{f}_i(\theta^{(k-1)}) \Big]
\end{aligned} \tag{48}$$

The above update rule will only be used for analysis. Considering the last term in the right hand side of (47); we have:

$$\begin{aligned}
\big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, w'_{ij} - w_i^* \big\rangle =& \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, w'_{ij} - \overline{w}_i^{(k)} \big\rangle + \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, \overline{w}_i^{(k)} - w_i^* \big\rangle \\
\overset{(a)}{\le}& \|\delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\| \cdot \|w'_{ij} - \overline{w}_i^{(k)}\| + \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, \overline{w}_i^{(k)} - w_i^* \big\rangle \\
\overset{(b)}{\le}& \alpha_w \|\delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\|^2 + \big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, \overline{w}_i^{(k)} - w_i^* \big\rangle
\end{aligned} \tag{49}$$

where (a) uses Cauchy-Schwarz inequality, and step (b) substitutes the proximal expressions of w'_{ij} in (38) and $\overline{w}_i^{(k)}$ in (49) followed by Lemma E.6. Averaging both sides of the inequality over all $0 \le j \le n_{Y_i} - 1$ and using the fact that the average of the second term on the right hand side of (49) is zero, we get:

$$\begin{aligned}
&\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \big[\big\langle \delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\,, w'_{ij} - w_i^* \big\rangle\big] \\
\le& \frac{\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \big[\|\delta_{ij}^w - \overline{f}_i(\theta^{(k-1)})\|^2\big] \\
=& \frac{\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \big[\|f_{ij}(\theta^{(k-1)}) - f_{ij}(\widetilde{\theta}) + \overline{f}_i(\widetilde{\theta}) - \overline{f}_i(\theta^{(k-1)})\|^2\big] \\
\overset{(a)}{=}& \frac{\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Big\|\Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\theta^{(k-1)}) \\
&\qquad - \Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\widetilde{\theta}) - \Big(\overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\widetilde{\theta}) - \frac{1}{n_{Y_i}} f_{ij}(\theta^{(k-1)}) + \frac{1}{n_{Y_i}} f_{ij}(\widetilde{\theta})\Big)\Big\|^2 \\
\overset{(b)}{\le}& \frac{2\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Bigg[\Big\|\Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\theta^{(k-1)}) - \Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\widetilde{\theta})\Big\|^2 \\
&\qquad\qquad + \Big\|\overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\widetilde{\theta}) - \frac{1}{n_{Y_i}} f_{ij}(\theta^{(k-1)}) + \frac{1}{n_{Y_i}} f_{ij}(\widetilde{\theta})\Big\|^2\Bigg] \\
=& \frac{2\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Bigg[\Big\|\Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\theta^{(k-1)}) - \Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\widetilde{\theta})\Big\|^2 \\
&\qquad\qquad + \Big\|\frac{n_{Y_i} - 1}{n_{Y_i}} \frac{1}{n_{Y_i} - 1} \sum_{\ell \ne j} \big[f_{i\ell}(\theta^{k-1}) - f_{i\ell}(\widetilde{\theta})\big]\Big\|^2\Bigg] \\
\overset{(c)}{\le}& \frac{2\alpha_w}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \Bigg[\Big\|\Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\theta^{(k-1)}) - \Big(1 - \frac{1}{n_{Y_i}}\Big) f_{ij}(\widetilde{\theta})\Big\|^2 \\
&\qquad\qquad + \Big(\frac{n_{Y_i} - 1}{n_{Y_i}}\Big)^2 \frac{1}{n_{Y_i} - 1} \sum_{\ell \ne j} \Big\|\big[f_{i\ell}(\theta^{k-1}) - f_{i\ell}(\widetilde{\theta})\big]\Big\|^2\Bigg] \\
\overset{(d)}{\le}& 2\alpha_w \Big[\Big(1 - \frac{1}{n_{Y_i}}\Big)^2 B_f^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2 + \frac{n_{Y_i} - 1}{n_{Y_i}^2} B_f^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2\Big] \\
=& 2\alpha_w B_f^2 \Big(1 - \frac{1}{n_{Y_i}}\Big) \|\theta^{(k-1)} - \widetilde{\theta}\|^2
\end{aligned} \tag{50}$$

where step (a) follows by adding and subtracting $\frac{1}{n_{Y_i}}\big(f_{ij}(\theta^{(k-1)}) - f_{ij}(\widetilde{\theta})\big)$, step (b) uses $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, step (c) applies Jensen's inequality to the second term, step (d) applies Assumption E.3 (uniformly bounded gradients implies the uniform Liphschitz continuity of the functions f_{ij}) to both terms. Averaging both sides of (47) over all $0 \le j \le n_{Y_i} - 1$, and using (50), we obtain:

$$
\begin{aligned}
\left(\frac{1}{2\alpha_w} + \gamma\right) \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \|w'_{ij} - w_i^*\|^2 \le{}& \frac{1}{2\alpha_w}\|w_i^{(k-1)} - w_i^*\|^2 - \frac{1}{2\alpha_w}\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \|w'_{ij} - w_i^{(k-1)}\|^2 \\
&+ \frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \left\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w'_{ij} - w_i^* \right\rangle + 2\alpha_w B_f^2 \Big(1 - \frac{1}{n_{Y_i}}\Big) \|\theta^{(k-1)} - \widetilde{\theta}\|^2
\end{aligned}
\tag{51}
$$

Finally, we relate the bound for w'_{ij} back to the bound for the dual variable $w_i^{(k)}$. Recall that, for each w_i, there is a probability $1/n_X$ that it will be selected and updated, and a probability of $(n_X - 1)/n_X$ that it will be kept the same as $w_i^{(k-1)}$. Furthermore, conditioned on the fact that w_i is selected, it will be updated to w'_{ij} with probability $1/n_{Y_i}$. Therefore, for each w_i, there is a probability $1/n_X n_{Y_i}$ that it will be updated to w'_{ij} for $j = 0, \ldots, n_{Y_i} - 1$, and a probability of $(n_X - 1)/n_X$ that it remains the same. Therefore, letting $\mathcal{F}_k$ denote the filtration of all events upto the beginning of iteration k (before the dual update step), we have:

$$
\begin{aligned}
\mathsf{E}\{w_i^{(k)} \mid \mathcal{F}_k\} &= \frac{1}{n_X}\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} w'_{ij} + \frac{n_X - 1}{n_X} w_i^{(k-1)} \\
\mathsf{E}\{w_i^{(k)} - w_i^* \mid \mathcal{F}_k\} &= \frac{1}{n_X}\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} (w'_{ij} - w_i^*) + \frac{n_X - 1}{n_X} (w_i^{(k-1)} - w_i^*) \\
\mathsf{E}\{\|w_i^{(k)} - w_i^*\|^2 \mid \mathcal{F}_k\} &= \frac{1}{n_X}\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \|w'_{ij} - w_i^*\|^2 + \frac{n_X - 1}{n_X} \|w_i^{(k-1)} - w_i^*\|^2 \\
\mathsf{E}\{\|w_i^{(k)} - w_i^{(k-1)}\|^2 \mid \mathcal{F}_k\} &= \frac{1}{n_X}\frac{1}{n_{Y_i}} \sum_{j=0}^{n_{Y_i}-1} \|w'_{ij} - w_i^{(k-1)}\|^2
\end{aligned}
\tag{52}
$$

Using (52) in (51), we obtain:

$$
\begin{aligned}
\left(\frac{1}{2\alpha_w} + \gamma\right) &\left(n_X \mathsf{E}\{\|w_i^{(k)} - w_i^*\|^2 \mid \mathcal{F}_k\} - (n_X - 1)\|w_i^{(k-1)} - w_i^*\|^2\right) \\
&\le \frac{1}{2\alpha_w}\|w_i^{(k-1)} - w_i^*\|^2 - \frac{n_X}{2\alpha_w}\mathsf{E}\{\|w_i^{(k)} - w_i^{(k-1)}\|^2 \mid \mathcal{F}_k\} \\
&\quad + 2\alpha_w B_f^2 \Big(1 - \frac{1}{n_{Y_i}}\Big)\|\theta^{(k-1)} - \widetilde{\theta}\|^2 \\
&\quad + n_X \mathsf{E}\Big\{\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w_i^{(k)} - w_i^* \rangle \mid \mathcal{F}_k\Big\} \\
&\quad - \big(n_X - 1\big)\Big\{\langle \overline{f}_i(\theta^{(k-1)}) - \overline{f}_i(\theta^*), w_i^{(k-1)} - w_i^* \rangle\Big\}
\end{aligned}
\tag{53}
$$

Summing both sides of (53) over $0 \le i \le n_X - 1$, using the fact that $\|w^{(k)} - w^*\|^2 = \sum_{i=0}^{n_X-1} \|w_i^{(k)} - w_i^*\|^2$, and then dividing by n_X, we get

$$
\begin{aligned}
&\left(\frac{1}{2\alpha_w} + \gamma\right)\left(\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} - \frac{n_X - 1}{n_X}\|w^{(k-1)} - w^*\|^2\right) \\
&\le \frac{1}{2\alpha_w n_X}\|w^{(k-1)} - w^*\|^2 - \frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 \mid \mathcal{F}_k\} \\
&+ 2\alpha_w B_f^2 \Big(1 - \overline{1/n_Y}\Big)\|\theta^{(k-1)} - \widetilde{\theta}\|^2 \\
&+ n_X \mathsf{E}\Big\{L(\theta^{(k-1)}, w^{(k)} - w^*) - L(\theta^*, w^{(k)} - w^*) \mid \mathcal{F}_k\Big\} \\
&- \big(n_X - 1\big)\Big\{L(\theta^{(k-1)}, w^{(k-1)} - w^*) - L(\theta^*, w^{(k-1)} - w^*)\Big\}
\end{aligned}
\tag{54}
$$

where, we have used the notation:

$$\overline{1/n_Y} := \frac{1}{n_X}\sum_{i=0}^{n_X-1} 1/n_{Y_i} \tag{55}$$

Rearranging and combining the common terms, we obtain the final dual bound:

$$\begin{aligned}
&\left(\frac{1}{2\alpha_w}+\gamma\right)\mathsf{E}\{\|w^{(k)}-w^*\|^2 \mid \mathcal{F}_k\} + \frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2 \mid \mathcal{F}_k\}\\
&\le \left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(k-1)}-w^*\|^2 + 2\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)\|\theta^{(k-1)}-\widetilde{\theta}\|^2\\
&+\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^*) - L(\theta^*,w^{(k)}-w^*) \mid \mathcal{F}_k\Big\}\\
&+\big(n_X-1\big)\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^{(k-1)}) - L(\theta^*,w^{(k)}-w^{(k-1)}) \mid \mathcal{F}_k\Big\}
\end{aligned} \tag{56}$$

Note that the above bound still have terms related to $L(\cdot,\cdot)$. We will combine these terms together with the L terms in the primal bound, and then bound them all together thereafter.

E.6 Primal Bound

Now we proceed to derive the bound for the primal variable θ. Specifically, we will focus on examining $\|\theta^{(k)}-\theta^*\|^2$, which can be written as

$$\begin{aligned}
\|\theta^{(k)}-\theta^*\|^2 &= \|\theta^{(k)}-\theta^{(k-1)}+\theta^{(k-1)}-\theta^*\|^2\\
&= \|\theta^{(k)}-\theta^{(k-1)}\|^2 + \|\theta^{(k-1)}-\theta^*\|^2 + 2\langle(\theta^{(k)}-\theta^{(k-1)}),(\theta^{(k-1)}-\theta^*)\rangle
\end{aligned} \tag{57}$$

Similar to the dual bound, we now bound the first term and the third term together. Introduce the following function of x (for fixed $\theta^{(k-1)}$ and $w^{(k)}$):

$$P_\theta(x) := \big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\, x\big\rangle + g(x) \tag{58}$$

The first part of the function is linear in x (and hence convex), and the second part of the function is μ-strongly convex (Assumption E.1). Recall the primal update rule in (31), which can be written in the following proximal mapping form:

$$\begin{aligned}
\theta^{(k)} &= \arg\min_\theta \Big\{\langle\delta_k^\theta,\theta\rangle + g(\theta) + \frac{1}{2\alpha_\theta}\|\theta-\theta^{(k-1)}\|^2\Big\}\\
&= \mathrm{prox}_{\alpha_\theta g}\Big\{\theta^{(k-1)}-\alpha_\theta\delta_k^\theta\Big\}
\end{aligned} \tag{59}$$

We now apply Lemma E.5 with $P(x)\equiv P_\theta(x)$ and the following identifications:

$f(x)=\big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\, x\big\rangle \quad g(x)=g(x) \quad v=\delta_k^\theta \quad x=\theta^{(k-1)} \quad x^{(+)}=\theta^{(k)} \quad y=\theta^* \quad \alpha=\alpha_\theta$

which leads to

$$\begin{aligned}
\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^*\Big\rangle + g(\theta^*) \ge&\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^{(k)}\Big\rangle + g(\theta^{(k)})\\
&+\frac{1}{\alpha_\theta}\langle(\theta^{(k-1)}-\theta^{(k)}),(\theta^*-\theta^{(k-1)})\rangle\\
&-\langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)},w^{(k)})),(\theta^{(k)}-\theta^*)\rangle\\
&+\frac{1}{\alpha_\theta}\|\theta^{(k-1)}-\theta^{(k)}\|^2 + \frac{\mu}{2}\|\theta^*-\theta^{(k)}\|^2
\end{aligned} \tag{60}$$

Rearranging the terms in the above inequality, we obtain

$$\begin{aligned}
&\|\theta^{(k-1)}-\theta^{(k)}\|^2 + \langle\theta^{(k)}-\theta^{(k-1)},\theta^{(k-1)}-\theta^*\rangle\\
\le&\alpha_\theta\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^*\rangle + \alpha_\theta g(\theta^*) - \alpha_\theta g(\theta^{(k)}) - \alpha_\theta\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^{(k)}\rangle
\end{aligned}$$

$$+ \alpha_\theta \langle \delta_k^\theta - L'(\theta^{(k-1)}, w^{(k)}), \theta^{(k)} - \theta^* \rangle - \frac{\alpha_\theta \mu}{2} \|\theta^* - \theta^{(k)}\|^2 \tag{61}$$

Using (61) to bound the first and the third term in (57), we obtain:

$$\begin{aligned}
\|\theta^{(k)} - \theta^*\|^2 &= \|\theta^{(k)} - \theta^{(k-1)}\|^2 + \|\theta^{(k-1)} - \theta^*\|^2 + 2\langle(\theta^{(k)} - \theta^{(k-1)}), (\theta^{(k-1)} - \theta^*)\rangle \\
&\overset{(a)}{=} \|\theta^{(k)} - \theta^{(k-1)}\|^2 - \|\theta^{(k-1)} - \theta^*\|^2 + 2\|\theta^{(k-1)} - \theta^*\|^2 \\
&\quad + 2\langle(\theta^{(k)} - \theta^{(k-1)}), (\theta^{(k-1)} - \theta^*)\rangle \\
&\leq \|\theta^{(k-1)} - \theta^*\|^2 - \|\theta^{(k)} - \theta^{(k-1)}\|^2 - \alpha_\theta \mu \|\theta^{(k)} - \theta^*\|^2 \\
&\quad + 2\alpha_\theta \Big[\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^* \Big\rangle + g(\theta^*) - \Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^{(k)} \Big\rangle - g(\theta^{(k)}) \Big] \\
&\quad + 2\alpha_\theta \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \theta^*)\rangle
\end{aligned} \tag{62}$$

where step (a) subtracts and adds the second term. Furthermore, note that θ^* is the optimal solution to the following optimization problem:

$$\theta^* = \arg\min_\theta \Big[L(\theta, w^*) + g(\theta) \Big]$$

which implies (from the fact that g is μ-strongly convex):

$$L(\theta^{(k)}, w^*) + g(\theta^{(k)}) \geq L(\theta^*, w^*) + g(\theta^*) + \frac{\mu}{2} \|\theta^{(k)} - \theta^*\|^2 \tag{63}$$

Multiplying both sides of the above inequality by $2\alpha_\theta$ and then adding it to (62), we obtain:

$$\begin{aligned}
(1 + 2\alpha_\theta \mu)\|\theta^{(k)} - \theta^*\|^2 &\leq \|\theta^{(k-1)} - \theta^*\|^2 - \|\theta^{(k)} - \theta^{(k-1)}\|^2 \\
&\quad + 2\alpha_\theta \Big[L(\theta^{(k)}, w^*) - L(\theta^*, w^*) \Big] \\
&\quad + 2\alpha_\theta \Big[\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^* \Big\rangle - \Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^{(k)} \Big\rangle \Big] \\
&\quad + 2\alpha_\theta \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \theta^*)\rangle
\end{aligned} \tag{64}$$

Next, we bound the last term in (64). To this end, we first introduce the following auxiliary variable $\overline{\theta}^{(k)}$, which is the updated primal variable if *full batch gradient* were used:

$$\begin{aligned}
\overline{\theta}^{(k)} &= \arg\min_\theta \Big[\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta \Big\rangle + g(\theta) + \frac{1}{2\alpha_\theta} \|\theta - \theta^{(k-1)}\|^2 \Big] \\
&= \mathrm{prox}_{\alpha_\theta g} \Big\{ \theta^{(k-1)} - \alpha_\theta L'_\theta(\theta^{(k-1)}, w^{(k)}) \Big\}
\end{aligned} \tag{65}$$

Note that both (59) and (65) are written in proximal mapping form. We now bound the last term (64):

$$\begin{aligned}
&\langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \theta^*)\rangle \\
\overset{(a)}{=} &\langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \overline{\theta}^{(k)})\rangle + \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\overline{\theta}^{(k)} - \theta^*)\rangle \\
\overset{(b)}{\leq} &\|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\| \cdot \|\theta^{(k)} - \overline{\theta}^{(k)}\| + \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\overline{\theta}^{(k)} - \theta^*)\rangle \\
\overset{(c)}{\leq} &\alpha_\theta \|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 + \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\overline{\theta}^{(k)} - \theta^*)\rangle
\end{aligned} \tag{66}$$

where step (a) adds and subtracts $\overline{\theta}^{(k)}$, step (b) uses Cauchy-Schwartz inequality, and step (c) substitutes (59) and (65) and then applies Lemma E.6. Let $\mathcal{F}_k^{(+)}$ denote the filtration of all events up to and including the dual update in the k-th iteration. Applying expectation to both sides of (66) conditioned on $\mathcal{F}_k^{(+)}$, we have:

$$\begin{aligned}
&\mathsf{E}\Big\{ \langle(\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \theta^*)\rangle \big| \mathcal{F}_k^{(+)} \Big\} \\
\leq &\alpha_\theta \mathsf{E}\Big\{ \|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 \big| \mathcal{F}_k^{(+)} \Big\} + \mathsf{E}\Big\{ \langle \delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)}), \overline{\theta}^{(k)} - \theta^* \rangle \Big| \mathcal{F}_k^{(+)} \Big\} \\
\overset{(a)}{=} &\alpha_\theta \mathsf{E}\Big\{ \|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 \big| \mathcal{F}_k^{(+)} \Big\} + \Big\langle \mathsf{E}\big[\delta_k^\theta \big| \mathcal{F}_k^{(+)}\big] - L'_\theta(\theta^{(k-1)}, w^{(k)}), \overline{\theta}^{(k)} - \theta^* \Big\rangle
\end{aligned}$$

$$\overset{(b)}{=}\alpha_\theta \mathsf{E}\Big\{\|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 \big| \mathcal{F}_k^{(+)}\Big\} \tag{67}$$

where step (a) uses the fact that $\overline{\theta}^{(k)}$, $\theta^{(k-1)}$ and $w^{(k)}$ are deterministic conditioned on $\mathcal{F}_k^{(+)}$, and step (b) uses the fact that the conditional expectation of δ_k^θ is the batch gradient.

We will now upper bound the right hand side of (67). To this end, we have:

$$\begin{aligned}
&\mathsf{E}\Big\{\|\delta_k^\theta - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 \big| \mathcal{F}_k^{(+)}\Big\} \\
\overset{(a)}{=}&\mathsf{E}\Big\{\|f'_{i_k j_k}\big(\theta^{(k-1)}\big)w_{i_k}^{(k)} - f'_{i_k j_k}\big(\widetilde{\theta}\big)w_{i_k}^{(k)} + L'_\theta(\widetilde{\theta}, w^{(k)}) - L'_\theta(\theta^{(k-1)}, w^{(k)})\|^2 \big| \mathcal{F}_k^{(+)}\Big\} \\
\overset{(b)}{\le}&\mathsf{E}\Big\{\|f'_{i_k j_k}\big(\theta^{(k-1)}\big)w_{i_k}^{(k)} - f'_{i_k j_k}\big(\widetilde{\theta}\big)w_{i_k}^{(k)}\|^2 \big| \mathcal{F}_k^{(+)}\Big\} \\
\overset{(c)}{\le}&B_w^2 \mathsf{E}\Big\{\|f'_{i_k j_k}\big(\theta^{(k-1)}\big) - f'_{i_k j_k}\big(\widetilde{\theta}\big)\|^2 \big| \mathcal{F}_k^{(+)}\Big\} \\
\overset{(d)}{\le}&B_w^2 B_\theta^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2
\end{aligned} \tag{68}$$

where step (a) uses the definition of δ_k^θ, step (b) uses $\mathsf{E}[X - \mathsf{E}[X]]^2 \le \mathsf{E}[X^2]$, step (c) uses Lemma E.7, and step (d) uses the Lipschitz continuity of the gradients (Assumption E.3). Substituting (68) into (67), we get:

$$\mathsf{E}\Big\{\langle(\delta^\theta_{i'_k j'_k} - L'_\theta(\theta^{(k-1)}, w^{(k)})), (\theta^{(k)} - \theta^*)\rangle \Big| \mathcal{F}_k^{(+)}\Big\} \le \alpha_\theta B_w^2 B_\theta^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2 \tag{69}$$

Finally, substituting (69) into (64) and then further applying expectation conditioned on $\mathcal{F}_k$, we obtain:

$$\begin{aligned}
(1 + 2\alpha_\theta\mu)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k\Big\} \le\, &\|\theta^{(k-1)} - \theta^*\|^2 - \mathsf{E}\Big\{\|\theta^{(k)} - \theta^{(k-1)}\|^2 \big| \mathcal{F}_k\Big\} \\
&+ 2\alpha_\theta \mathsf{E}\Big\{\Big[L(\theta^{(k)}, w^*) - L(\theta^*, w^*)\Big] \big| \mathcal{F}_k\Big\} \\
&+ 2\alpha_\theta \mathsf{E}\Big\{\Big[\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^*\Big\rangle - \Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^{(k)}\Big\rangle\Big] \big| \mathcal{F}_k\Big\} \\
&+ 2\alpha_\theta^2 B_w^2 B_\theta^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2
\end{aligned} \tag{70}$$

Dividing both sides by $2\alpha_\theta$ and combining common terms, we obtain the final bound for the primal variable:

$$\begin{aligned}
&\Big(\frac{1}{2\alpha_\theta} + \mu\Big)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k\Big\} + \frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)} - \theta^{(k-1)}\|^2 \big| \mathcal{F}_k\Big\} \\
\le&\frac{1}{2\alpha_\theta}\|\theta^{(k-1)} - \theta^*\|^2 + \alpha_\theta B_w^2 B_\theta^2 \|\theta^{(k-1)} - \widetilde{\theta}\|^2 + \mathsf{E}\Big\{\Big[L(\theta^{(k)}, w^*) - L(\theta^*, w^*)\Big] \big| \mathcal{F}_k\Big\} \\
&+ \mathsf{E}\Big\{\Big[\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^*\Big\rangle - \Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^{(k)}\Big\rangle\Big] \big| \mathcal{F}_k\Big\}
\end{aligned} \tag{71}$$

E.7 Convergence for Option I

Based on the derived primal and dual bounds above, we now proceed to prove the convergence of SVRPDA-I with Option I: updating $\tilde{\theta}$ using the most recent $\theta^{(k)}$ (see Algorithm 1).

Adding (56) and (71) we obtain the total bound for the primal and dual variable updates:

$$
\begin{aligned}
&\left(\frac{1}{2\alpha_w}+\gamma\right)\mathsf{E}\{\|w^{(k)}-w^*\|^2 \mid \mathcal{F}_k\}+\frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2 \mid \mathcal{F}_k\}\\
&\left(\frac{1}{2\alpha_\theta}+\mu\right)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\Big|\mathcal{F}_k\Big\}+\frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\}\\
&\le\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(k-1)}-w^*\|^2+\frac{1}{2\alpha_\theta}\|\theta^{(k-1)}-\theta^*\|^2\\
&+\Big(2\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+\alpha_\theta B_w^2 B_\theta^2\Big)\|\theta^{(k-1)}-\widetilde{\theta}\|^2\\
&+\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^*)-L(\theta^*,w^{(k)}-w^*)\mid\mathcal{F}_k\Big\}\\
&+\mathsf{E}\Big\{\Big[L(\theta^{(k)},w^*)-L(\theta^*,w^*)\Big]\Big|\mathcal{F}_k\Big\}+\mathsf{E}\Big\{\Big[\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^*\Big\rangle-\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^{(k)}\Big\rangle\Big]\Big|\mathcal{F}_k\Big\}\\
&+\big(n_X-1\big)\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\mid\mathcal{F}_k\Big\}
\end{aligned}
\tag{72}
$$

Next, we need to upper bound the L terms on the right-hand side of the above inequality. To this end, we first show the following inequality:

$$
\begin{aligned}
&-L(\theta^*,w^{(k)}-w^*)+L(\theta^{(k)},w^*)-L(\theta^*,w^*)+\big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^*\big\rangle\\
&-\big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^{(k)}\big\rangle+L(\theta^{(k)},w^{(k)}-w^*)\\
&=-L(\theta^*,w^{(k)})+L(\theta^{(k)},w^{(k)})+\big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^*-\theta^{(k)}\big\rangle\big\rangle\\
&\overset{(a)}{\le}-\big\langle L'_\theta(\theta^{(k)},w^{(k)}),\theta^*-\theta^{(k)}\big\rangle+\big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^*-\theta^{(k)}\big\rangle\\
&=\big\langle L'_\theta(\theta^{(k)},w^{(k)})-L'_\theta(\theta^{(k-1)},w^{(k)}),\theta^{(k)}-\theta^*\big\rangle\\
&\overset{(b)}{=}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\Big\langle\Big(\overline{f}'_i(\theta^{(k)})-\overline{f}'_i(\theta^{(k-1)})\Big)w_i^{(k)},\theta^{(k)}-\theta^*\Big\rangle\\
&\le\left|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\Big\langle\Big(\overline{f}'_i(\theta^{(k)})-\overline{f}'_i(\theta^{(k-1)})\Big)w_i^{(k)},\theta^{(k)}-\theta^*\Big\rangle\right|\\
&\overset{(c)}{\le}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\left|\Big\langle\Big(\overline{f}'_i(\theta^{(k)})-\overline{f}'_i(\theta^{(k-1)})\Big)w_i^{(k)},\theta^{(k)}-\theta^*\Big\rangle\right|\\
&\overset{(d)}{\le}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\left\|\Big(\overline{f}'_i(\theta^{(k)})-\overline{f}'_i(\theta^{(k-1)})\Big)w_i^{(k)}\right\|\cdot\|\theta^{(k)}-\theta^*\|\\
&\le\frac{1}{n_X}\sum_{i=0}^{n_X-1}\|\overline{f}'_i(\theta^{(k)})-\overline{f}'_i(\theta^{(k-1)})\|\cdot\|w_i^{(k)}\|\cdot\|\theta^{(k)}-\theta^*\|\\
&\overset{(e)}{=}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\left\|\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1}\Big(f'_{\theta^{(k)}}(x_i,y_{ij})-f'_{\theta^{(k-1)}}(x_i,y_{ij})\Big)\right\|\cdot\|w_i^{(k)}\|\cdot\|\theta^{(k)}-\theta^*\|\\
&\overset{(f)}{\le}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1}\left\|f'_{\theta^{(k)}}(x_i,y_{ij})-f'_{\theta^{(k-1)}}(x_i,y_{ij})\right\|\cdot\|w_i^{(k)}\|\cdot\|\theta^{(k)}-\theta^*\|\\
&\overset{(g)}{\le}\frac{1}{n_X}\sum_{i=0}^{n_X-1}\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1}B_\theta B_w\|\theta^{(k)}-\theta^{(k-1)}\|\cdot\|\theta^{(k)}-\theta^*\|\\
&=B_\theta B_w\|\theta^{(k)}-\theta^{(k-1)}\|\cdot\|\theta^{(k)}-\theta^*\|
\end{aligned}
$$

$$\overset{(h)}{\leq} \frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \tag{73}$$

where step (a) uses the fact that $L(\theta, w)$ is convex with respect to the θ, step (b) substitutes the expression of L'_θ, step (c) uses Jensen's inequality, step (d) uses Cauchy-Schwartz inequality, step (e) substitutes the expression for $\overline{f}'_i$, step (f) uses Jensen's inequality, step (g) uses the Lipschitz gradient property of f'_θ and Lemma E.7, step (h) uses $ab \leq \frac{1}{\beta_0}a^2+\beta_0 b^2$. In consequence, the above inequality implies that

$$\begin{aligned}
&-L(\theta^*, w^{(k)}-w^*)+L(\theta^{(k)}, w^*)-L(\theta^*, w^*)+\left\langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^*\right\rangle-\left\langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^{(k)}\right\rangle \\
\leq &-L(\theta^{(k)}, w^{(k)}-w^*)+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2
\end{aligned} \tag{74}$$

Using (74), the L terms in (72) becomes (notice that we keep the first and last L terms in (72) intact)

$$\begin{aligned}
&L(\theta^{(k-1)}, w^{(k)}-w^*)-L(\theta^*, w^{(k)}-w^*)+L(\theta^{(k)}, w^*)-L(\theta^*, w^*)+\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^*\Big\rangle \\
&\quad-\Big\langle L'_\theta(\theta^{(k-1)}, w^{(k)}),\, \theta^{(k)}\Big\rangle+(n_X-1)\Big[L(\theta^{(k-1)}, w^{(k)}-w^{(k-1)})-L(\theta^*, w^{(k)}-w^{(k-1)})\Big] \\
\overset{(a)}{\leq}&L(\theta^{(k-1)}, w^{(k)}-w^*)-L(\theta^{(k)}, w^{(k)}-w^*)+(n_X-1)\Big[L(\theta^{(k-1)}, w^{(k)}-w^{(k-1)})-L(\theta^*, w^{(k)}-w^{(k-1)})\Big] \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
=&L\Big(\theta^{(k-1)}, w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big)-L\Big(\theta^{(k)}, w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big) \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
\overset{(b)}{\leq}&\Big\langle L'_\theta\Big(\theta^{(k-1)}, w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big), \theta^{(k-1)}-\theta^{(k)}\Big\rangle \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
=&\Big\langle\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k)}-w_i^*+(n_X-1)(w_i^{(k)}-w_i^{(k-1)})\Big), \theta^{(k-1)}-\theta^{(k)}\Big\rangle \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
=&\Big\langle\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}+n_X(w_i^{(k)}-w_i^{(k-1)})-w_i^*\Big), \theta^{(k-1)}-\theta^{(k)}\Big\rangle \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
\overset{(c)}{\leq}&\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}+n_X(w_i^{(k)}-w_i^{(k-1)})-w_i^*\Big)\Big\|^2+\frac{1}{\beta_1}\Big\|\theta^{(k-1)}-\theta^{(k)}\Big\|^2 \\
&\quad+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
=&\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}-w_i^*\Big)+\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2 \\
&\quad+\frac{1}{\beta_1}\Big\|\theta^{(k-1)}-\theta^{(k)}\Big\|^2+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2 \\
\overset{(d)}{\leq}&2\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}-w_i^*\Big)\Big\|^2+2\beta_1\Big\|\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2 \\
&\quad+\frac{1}{\beta_1}\Big\|\theta^{(k-1)}-\theta^{(k)}\Big\|^2+\frac{B_\theta B_w}{\beta_0}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\beta_0 B_\theta B_w\|\theta^{(k)}-\theta^*\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(e)}{\le} \frac{2\beta_1 B_f^2}{n_X} \sum_{i=0}^{n_X-1} \|w_i^{(k-1)} - w_i^*\|^2 + 2\beta_1 \Big\| \sum_{i=0}^{n_X-1} \overline{f}_i'(\theta^{(k-1)})(w_i^{(k)} - w_i^{(k-1)}) \Big\|^2 \\
&\quad + \Big(\frac{1}{\beta_1} + \frac{B_\theta B_w}{\beta_0}\Big) \big\|\theta^{(k-1)} - \theta^{(k)}\big\|^2 + \beta_0 B_\theta B_w \|\theta^{(k)} - \theta^*\|^2
\end{aligned}
\tag{75}
$$

where step (a) applies (74), step (b) uses convexity of L in θ, step (c) uses $\langle a, b\rangle \le \beta_1\|a\|^2 + \frac{1}{\beta_1}\|b\|^2$, for some $\beta_1 > 0$ to be chosen later, step (d) uses $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and step (e) applies Jensen's inequality and bounded gradient assumption to the first term. Before we proceed, we note that, by taking expectation of the second term conditioned on $\mathcal{F}_k$, we get

$$
\begin{aligned}
&\mathsf{E}\Big\{ \Big\| \sum_{i=0}^{n_X-1} \overline{f}_i'(\theta^{(k-1)})(w_i^{(k)} - w_i^{(k-1)}) \Big\|^2 \Big| \mathcal{F}_k \Big\} \\
=& \sum_{i=0}^{n_X-1} \sum_{j=0}^{n_{Y_i}-1} \frac{1}{n_X n_{Y_i}} \Big\| \overline{f}_i'(\theta^{(k-1)})(w_{ij}' - w_i^{(k-1)}) \Big\|^2 \\
\le& B_f^2 \sum_{i=0}^{n_X-1} \sum_{j=0}^{n_{Y_i}-1} \frac{1}{n_X n_{Y_i}} \Big\| w_{ij}' - w_i^{(k-1)} \Big\|^2 \\
=& B_f^2 \sum_{i=0}^{n_X-1} \mathsf{E}\big\{ \|w_i^{(k)} - w_i^{(k-1)}\|^2 \big| \mathcal{F}_k \big\}
\end{aligned}
\tag{76}
$$

Therefore, the conditional expectation of all the L terms are bounded by

$$
\begin{aligned}
&\frac{2\beta_1 B_f^2}{n_X} \sum_{i=0}^{n_X-1} \|w_i^{(k-1)} - w_i^*\|^2 + 2\beta_1 B_f^2 \sum_{i=0}^{n_X-1} \mathsf{E}\{\|w_i^{(k)} - w_i^{(k-1)}\|^2 | \mathcal{F}_k\} \\
&+ \Big(\frac{1}{\beta_1} + \frac{B_\theta B_w}{\beta}\Big) \mathsf{E}\big\{ \big\|\theta^{(k-1)} - \theta^{(k)}\big\|^2 \big| \mathcal{F}_k \big\} + \beta_0 B_\theta B_w \mathsf{E}\big\{ \|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k \big\} \\
=& \frac{2\beta_1 B_f^2}{n_X} \|w^{(k-1)} - w^*\|^2 + 2\beta_1 B_f^2 \mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 | \mathcal{F}_k\} \\
&+ \Big(\frac{1}{\beta_1} + \frac{B_\theta B_w}{\beta_0}\Big) \mathsf{E}\big\{ \big\|\theta^{(k-1)} - \theta^{(k)}\big\|^2 \big| \mathcal{F}_k \big\} + \beta_0 B_\theta B_w \mathsf{E}\big\{ \|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k \big\}
\end{aligned}
\tag{77}
$$

Therefore, the total bound (72) becomes

$$
\begin{aligned}
&\left(\frac{1}{2\alpha_w} + \gamma \right) \mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} + \frac{1}{2\alpha_w} \mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 \mid \mathcal{F}_k\} \\
&+ \left(\frac{1}{2\alpha_\theta} + \mu \right) \mathsf{E}\Big\{ \|\theta^{(k)} - \theta^*\|^2 \Big| \mathcal{F}_k \Big\} + \frac{1}{2\alpha_\theta} \mathsf{E}\Big\{ \|\theta^{(k)} - \theta^{(k-1)}\|^2 \Big| \mathcal{F}_k \Big\} \\
&\le \left(\frac{1}{2\alpha_w} + \frac{\gamma(n_X - 1)}{n_X} \right) \|w^{(k-1)} - w^*\|^2 + \frac{1}{2\alpha_\theta} \|\theta^{(k-1)} - \theta^*\|^2 \\
&+ \Big(2\alpha_w B_f^2 \Big(1 - \overline{1/n_Y}\Big) + \alpha_\theta B_w^2 B_\theta^2 \Big) \|\theta^{(k-1)} - \widetilde{\theta}\|^2 \\
&+ \frac{2\beta_1 B_f^2}{n_X} \|w^{(k-1)} - w^*\|^2 + 2\beta_1 B_f^2 \mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 | \mathcal{F}_k\} \\
&+ \Big(\frac{1}{\beta_1} + \frac{B_\theta B_w}{\beta_0}\Big) \mathsf{E}\big\{ \big\|\theta^{(k-1)} - \theta^{(k)}\big\|^2 \big| \mathcal{F}_k \big\} + \beta_0 B_\theta B_w \mathsf{E}\big\{ \|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k \big\}
\end{aligned}
\tag{78}
$$

By combining the common terms, we obtain

$$
\left(\frac{1}{2\alpha_\theta} + \mu - \beta_0 B_\theta B_w \right) \mathsf{E}\Big\{ \|\theta^{(k)} - \theta^*\|^2 \Big| \mathcal{F}_k \Big\} + \left(\frac{1}{2\alpha_w} + \gamma \right) \mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\}
$$

$$
\begin{aligned}
&+\Big(\frac{1}{2\alpha_\theta}-\frac{1}{\beta_1}-\frac{B_\theta B_w}{\beta_0}\Big)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\big|\mathcal{F}_k\Big\}+\Big(\frac{1}{2\alpha_w}-2\beta_1 B_f^2\Big)\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2\mid\mathcal{F}_k\}\\
\le&\frac{1}{2\alpha_\theta}\|\theta^{(k-1)}-\theta^*\|^2+\Big(\frac{1}{2\alpha_w}+\gamma-\frac{\gamma}{n_X}+\frac{2\beta_1 B_f^2}{n_X}\Big)\|w^{(k-1)}-w^*\|^2\\
&+\Big(2\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+\alpha_\theta B_w^2 B_\theta^2\Big)\|\theta^{(k-1)}-\widetilde{\theta}\|^2
\end{aligned}
\tag{79}
$$

Applying inequality $\|x+y\|^2\le 2\|x\|^2+2\|y\|^2$ to the last term in (79), we obtain

$$
\begin{aligned}
&\Bigg(\frac{1}{2\alpha_\theta}+\mu-\beta_0 B_\theta B_w\Bigg)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\big|\mathcal{F}_k\Big\}+\Bigg(\frac{1}{2\alpha_w}+\gamma\Bigg)\mathsf{E}\{\|w^{(k)}-w^*\|^2\mid\mathcal{F}_k\}\\
&+\Big(\frac{1}{2\alpha_\theta}-\frac{1}{\beta_1}-\frac{B_\theta B_w}{\beta_0}\Big)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\big|\mathcal{F}_k\Big\}+\Big(\frac{1}{2\alpha_w}-2\beta_1 B_f^2\Big)\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2\mid\mathcal{F}_k\}\\
\le&\Big(\frac{1}{2\alpha_\theta}+4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\|\theta^{(k-1)}-\theta^*\|^2\\
&+\Big(\frac{1}{2\alpha_w}+\gamma-\frac{\gamma}{n_X}+\frac{2\beta_1 B_f^2}{n_X}\Big)\|w^{(k-1)}-w^*\|^2+\Big(4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}
\tag{80}
$$

Taking full expectation of the above inequality, we obtain:

$$
\begin{aligned}
&\Big(\frac{1}{2\alpha_\theta}+\mu-\beta_0 B_\theta B_w\Big)\mathsf{E}\|\theta^{(k)}-\theta^*\|^2+\Big(\frac{1}{2\alpha_w}+\gamma\Big)\mathsf{E}\|w^{(k)}-w^*\|^2\\
&+\Big(\frac{1}{2\alpha_\theta}-\frac{1}{\beta_1}-\frac{B_\theta B_w}{\beta_0}\Big)\mathsf{E}\|\theta^{(k)}-\theta^{(k-1)}\|^2+\Big(\frac{1}{2\alpha_w}-2\beta_1 B_f^2\Big)\mathsf{E}\|w^{(k)}-w^{(k-1)}\|^2\\
\le&\Big(\frac{1}{2\alpha_\theta}+4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\mathsf{E}\|\theta^{(k-1)}-\theta^*\|^2\\
&+\Big(\frac{1}{2\alpha_w}+\gamma-\frac{\gamma}{n_X}+\frac{2\beta_1 B_f^2}{n_X}\Big)\mathsf{E}\|w^{(k-1)}-w^*\|^2\\
&+\Big(4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\mathsf{E}\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}
\tag{81}
$$

In order for the above inequality to converge, the hyperparameters need to be chosen to satisfy the following conditions:

$$
\begin{aligned}
\frac{1}{2\alpha_\theta}&\ge\frac{1}{\beta_1}+\frac{B_\theta B_w}{\beta_0}\\
\alpha_w&\le\frac{1}{4B_f^2\beta_1}\\
\beta_1&<\frac{\gamma}{2B_f^2}\\
4\alpha_w B_f^2(1-\overline{1/n_Y})+2\alpha_\theta B_w^2 B_\theta^2&<\mu-\beta_0 B_\theta B_w
\end{aligned}
\tag{82}
$$

which simplifies the recursion to be

$$
\begin{aligned}
&\Big(\frac{1}{2\alpha_\theta}+\mu-\beta_0 B_\theta B_w\Big)\mathsf{E}\|\theta^{(k)}-\theta^*\|^2+\Big(\frac{1}{2\alpha_w}+\gamma\Big)\mathsf{E}\|w^{(k)}-w^*\|^2\\
\le&\Big(\frac{1}{2\alpha_\theta}+4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\mathsf{E}\|\theta^{(k-1)}-\theta^*\|^2\\
&+\Big(\frac{1}{2\alpha_w}+\gamma-\frac{\gamma}{n_X}+\frac{2\beta_1 B_f^2}{n_X}\Big)\mathsf{E}\|w^{(k-1)}-w^*\|^2\\
&+\Big(4\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2 B_\theta^2\Big)\mathsf{E}\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}
\tag{83}
$$

Inequality (83) can also be further written as

$$\begin{aligned}
&\mathsf{E}\|\theta^{(k)}-\theta^*\|^2+\frac{\frac{1}{2\alpha_w}+\gamma}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}\mathsf{E}\|w^{(k)}-w^*\|^2\\
\le& r_P\mathsf{E}\|\theta^{(k-1)}-\theta^*\|^2+r_D\cdot\frac{\frac{1}{2\alpha_w}+\gamma}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}\mathsf{E}\|w^{(k-1)}-w^*\|^2\\
&+\frac{4\alpha_wB_f^2\big(1-\overline{1/n_Y}\big)+2\alpha_\theta B_w^2B_\theta^2}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}\mathsf{E}\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}\tag{84}$$

where r_P and r_D are the primal and the dual ratios, defined as

$$\begin{aligned}
r_P&=\frac{\frac{1}{2\alpha_\theta}+4\alpha_wB_f^2\Big(1-\overline{1/n_Y}\Big)+2\alpha_\theta B_w^2B_\theta^2}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}\\
r_D&=1-\frac{1}{n_X}\frac{2\alpha_w(\gamma-2\beta_1B_f^2)}{1+2\alpha_w\gamma}
\end{aligned}$$

We choose β_0, β_1, and the primal and the dual step-sizes to be

$$\begin{aligned}
\beta_0&=\frac{\mu}{2B_\theta B_w},\quad \beta_1=\frac{\gamma}{4B_f^2}\\
\alpha_\theta&=\frac{\frac{1}{\mu}}{64n_X\big(\frac{B_f^2}{\mu\gamma}+\frac{B_\theta^2B_w^2}{\mu^2}\big)+n_X}=\frac{1}{\mu}\cdot\frac{1}{64n_X\kappa+n_X}\\
\alpha_w&=\frac{\frac{1}{\gamma}}{64\big(\frac{B_f^2}{\mu\gamma}+\frac{B_\theta^2B_w^2}{\mu^2}\big)+1}=\frac{1}{\gamma}\cdot\frac{1}{64\kappa+1}
\end{aligned}$$

where κ is the condition number defined as

$$\kappa=\frac{B_f^2}{\mu\gamma}+\frac{B_\theta^2B_w^2}{\mu^2}\tag{85}$$

It can be verified that the above choice of step-sizes satisfies the condition (82). With our choice of the parameters, we also have

$$\frac{\frac{1}{2\alpha_w}+\gamma}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}=\frac{\frac{1}{2\alpha_w}+\gamma}{\frac{1}{2\alpha_\theta}+\frac{\mu}{2}}=\frac{\gamma}{\mu}\cdot\frac{64\kappa+3}{64n_X\kappa+n_X+1}\tag{86}$$

and

$$\begin{aligned}
\frac{4\alpha_wB_f^2\big(1-\overline{1/n_Y}\big)+2\alpha_\theta B_w^2B_\theta^2}{\frac{1}{2\alpha_\theta}+\mu-\beta_0B_\theta B_w}&=\frac{4\alpha_wB_f^2\big(1-\overline{1/n_Y}\big)+2\alpha_\theta B_w^2B_\theta^2}{\frac{1}{2\alpha_\theta}+\frac{\mu}{2}}\\
&=\frac{\frac{8B_f^2\big(1-\overline{1/n_Y}\big)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{(64\kappa+1)(64n_X\kappa+n_X+1)}
\end{aligned}\tag{87}$$

Substituting (86)–(87) into (84), we obtain

$$\begin{aligned}
&\mathsf{E}\|\theta^{(k)}-\theta^*\|^2+\frac{\gamma}{\mu}\cdot\frac{64\kappa+3}{64n_X\kappa+n_X+1}\mathsf{E}\|w^{(k)}-w^*\|^2\\
\le& r_P\mathsf{E}\|\theta^{(k-1)}-\theta^*\|^2+r_D\cdot\frac{\gamma}{\mu}\cdot\frac{64\kappa+3}{64n_X\kappa+n_X+1}\mathsf{E}\|w^{(k-1)}-w^*\|^2\\
&+\frac{\frac{8B_f^2\big(1-\overline{1/n_Y}\big)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{(64\kappa+1)(64n_X\kappa+n_X+1)}\mathsf{E}\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}\tag{88}$$

Furthermore, the primal and the dual ratios can be upper bounded as

$$\begin{aligned}
r_P &= \frac{1 + 8\alpha_\theta \alpha_w B_f^2 \Big(1 - \overline{1/n_Y}\Big) + 4\alpha_\theta^2 B_w^2 B_\theta^2}{1 + 2\alpha_\theta \mu - 2\alpha_\theta \beta_0 B_\theta B_w} \\
&= \frac{1 + 8\alpha_\theta \alpha_w B_f^2 \Big(1 - \overline{1/n_Y}\Big) + 4\alpha_\theta^2 B_w^2 B_\theta^2}{1 + \alpha_\theta \mu} \\
&= 1 - \frac{\alpha_\theta \mu - 8\alpha_\theta \alpha_w B_f^2 \Big(1 - \overline{1/n_Y}\Big) - 4\alpha_\theta^2 B_w^2 B_\theta^2}{1 + \alpha_\theta \mu} \\
&\leq 1 - \frac{1}{\frac{1024}{13} n_X \kappa + \frac{16}{13} n_X + \frac{16}{13}} \\
&\leq 1 - \frac{1}{78.8 n_X \kappa + 1.3 n_X + 1.3} \qquad (89) \\
r_D &= 1 - \frac{1}{n_X} \frac{2\alpha_w(\gamma - 2\beta_1 B_f^2)}{1 + 2\alpha_w \gamma} \\
&= 1 - \frac{1}{n_X} \frac{\alpha_w \gamma}{1 + 2\alpha_w \gamma} \\
&= 1 - \frac{1}{64 n_X \kappa + 3 n_X} < r_P \qquad (90)
\end{aligned}$$

Therefore, inequality (88) can be further upper bounded as

$$\begin{aligned}
&\mathsf{E}\|\theta^{(k)} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|w^{(k)} - w^*\|^2 \\
\leq & r_P \Big(\mathsf{E}\|\theta^{(k-1)} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|w^{(k-1)} - w^*\|^2 \Big) \\
& + \frac{\frac{8B_f^2 \left(1 - \overline{1/n_Y}\right)}{\mu\gamma} + \frac{4B_w^2 B_\theta^2}{\mu^2} \cdot \frac{1}{n_X}}{(64\kappa + 1)(64 n_X \kappa + n_X + 1)} \mathsf{E}\|\widetilde{\theta} - \theta^*\|^2 \\
\leq & r_P \Big(\mathsf{E}\|\theta^{(k-1)} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|w^{(k-1)} - w^*\|^2 \Big) \\
& + \frac{\frac{8B_f^2 \left(1 - \overline{1/n_Y}\right)}{\mu\gamma} + \frac{4B_w^2 B_\theta^2}{\mu^2} \cdot \frac{1}{n_X}}{(64\kappa + 1)(64 n_X \kappa + n_X + 1)} \Big(\mathsf{E}\|\widetilde{\theta} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|\tilde{w} - w^*\|^2 \Big) \\
\overset{(a)}{=} & r_P \Big(\mathsf{E}\|\theta^{(k-1)} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|w^{(k-1)} - w^*\|^2 \Big) \\
& + \frac{\frac{8B_f^2 \left(1 - \overline{1/n_Y}\right)}{\mu\gamma} + \frac{4B_w^2 B_\theta^2}{\mu^2} \cdot \frac{1}{n_X}}{(64\kappa + 1)(64 n_X \kappa + n_X + 1)} \Big(\mathsf{E}\|\widetilde{\theta}_{s-1} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|\tilde{w}_{s-1} - w^*\|^2 \Big)
\end{aligned} \tag{91}$$

where step (a) uses the fact that the $\widetilde{\theta} = \widetilde{\theta}_{s-1}$ and $\tilde{w} = \tilde{w}_{s-1}$ when we are considering the s-th stage/outer-loop (see Algorithm 1 in the main paper). Define the following Lyapunov functions:

$$\begin{aligned}
P_{s,k} &= \mathsf{E}\|\theta^{(k)} - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|w^{(k)} - w^*\|^2 \\
P_s &= \mathsf{E}\|\widetilde{\theta}_s - \theta^*\|^2 + \frac{\gamma}{\mu} \cdot \frac{64\kappa + 3}{64 n_X \kappa + n_X + 1} \mathsf{E}\|\tilde{w}_s - w^*\|^2
\end{aligned}$$

As a result, we can rewrite inequality (91) as

$$P_{s,k} \leq r_P \cdot P_{s,k-1} + \frac{\frac{8B_f^2 \left(1 - \overline{1/n_Y}\right)}{\mu\gamma} + \frac{4B_w^2 B_\theta^2}{\mu^2} \cdot \frac{1}{n_X}}{(64\kappa + 1)(64 n_X \kappa + n_X + 1)} P_{s-1}$$

$$\le \left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)\cdot P_{s,k-1}+\frac{\frac{8B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{(64\kappa+1)(64n_X\kappa+n_X+1)}P_{s-1} \tag{92}$$

Furthermore, at the s-th stage (outer loop iteration), when Option I is used in Algorithm 1 in the main paper, we have $\widetilde{\theta}_s=\theta^{(M)}$ and $\tilde{w}_s=w^{(M)}$. Therefore, it holds that

$$\begin{aligned}
&P_s=P_{s,M}\\
&\overset{(a)}{\le}\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M P_{s,0}\\
&\quad+\sum_{k=0}^{M-1}\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^k\times\frac{\frac{8B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{(64\kappa+1)(64n_X\kappa+n_X+1)}P_{s-1}\\
&\le\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M P_{s,0}\\
&\quad+\sum_{k=0}^{+\infty}\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^k\times\frac{\frac{8B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{(64\kappa+1)(64n_X\kappa+n_X+1)}P_{s-1}\\
&=\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M P_{s,0}\\
&\quad+\left(\frac{8B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}\right)\frac{78.8n_X\kappa+1.3n_X+1.3}{(64\kappa+1)(64n_X\kappa+n_X+1)}P_{s-1}\\
&\le\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M P_{s,0}+\left(\frac{8B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{4B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}\right)\frac{1.3}{64\kappa+1}P_{s-1}\\
&\le\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M P_{s,0}+\frac{\frac{16B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{8B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{64\kappa+1}P_{s-1}\\
&\overset{(b)}{=}\left[\left(1-\frac{1}{78.8n_X\kappa+1.3n_X+1.3}\right)^M+\frac{\frac{16B_f^2\left(1-\overline{1/n_Y}\right)}{\mu\gamma}+\frac{8B_w^2B_\theta^2}{\mu^2}\cdot\frac{1}{n_X}}{64\kappa+1}\right]P_{s-1}\\
&\le\left[e^{-\frac{M}{78.8n_X\kappa+1.3n_X+1.3}}+\frac{1}{4}\right]P_{s-1}
\end{aligned}\tag{93}$$

where step (a) iteratively applies inequality (91), and step (b) uses the fact $P_{s,0}=P_{s-1}$ (because $\theta^{(0)}=\widetilde{\theta}_{s-1}$ and $w^{(0)}=\tilde{w}_{s-1}$ as shown in Algorithm 1). Choosing $M=\lceil 78.8n_X\kappa+1.3n_X+1.3\rceil$, where $\lceil\cdot\rceil$ denotes the roundup operation, we have

$$P_s\le(e^{-1}+1/4)P_{s-1}<\frac{3}{4}P_{s-1}\le\left(\frac{3}{4}\right)^s P_0 \tag{94}$$

Therefore, P_s converges to zero at a linear rate of $3/4$. Furthermore, we requires a total of $\ln\frac{1}{\epsilon}$ outer loop iterations to reach ϵ-solution. And for each outer loop iteration, it requires M steps of inner-loop primal-dual updates, which is $O(1)$ per step (in number oracle calls), and $O(n_Xn_Y)$ for evaluating the batch gradients for the control variates, where $n_Y=(n_{Y_0}+\cdots+n_{Y_{n_X-1}})/n_X$. Therefore, the complexity per outer loop iteration is $O(n_Xn_Y+M)$, and the total complexity is

$$O\Big(\big(n_Xn_Y+n_X\kappa+n_X\big)\ln\frac{1}{\epsilon}\Big) \tag{95}$$

Noting that $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2\le P_s$, the above bound (94) also implies that $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2$ also converges to zero at a linear rate of $3/4$ and the total complexity to reach $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2\le\epsilon$ is also given by (95).

E.8 Convergence for Option II

Next, we move on to analyze the Option II case of the algorithm, wherein $\widetilde{\theta}$ at the end of each state is chosen to be one of the M previous $\theta^{(k)}$ values (see Algorithm 1 in the main paper).

Adding (56) and (71) we obtain the total bound for the primal and dual variable updates:

$$
\begin{aligned}
&\left(\frac{1}{2\alpha_w}+\gamma\right)\mathsf{E}\{\|w^{(k)}-w^*\|^2 \mid \mathcal{F}_k\}+\frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2 \mid \mathcal{F}_k\}\\
&\left(\frac{1}{2\alpha_\theta}+\mu\right)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\Big|\mathcal{F}_k\Big\}+\frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\}\\
&\leq\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(k-1)}-w^*\|^2+\frac{1}{2\alpha_\theta}\|\theta^{(k-1)}-\theta^*\|^2\\
&+\Big(2\alpha_w B_f^2\Big(1-\overline{1/n_Y}\Big)+\alpha_\theta B_w^2 B_\theta^2\Big)\|\theta^{(k-1)}-\widetilde{\theta}\|^2\\
&+\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^*)-L(\theta^*,w^{(k)}-w^*)\mid\mathcal{F}_k\Big\}\\
&+\big(n_X-1)\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\mid\mathcal{F}_k\Big\}\\
&+\mathsf{E}\Big\{\Big[L(\theta^{(k)},w^*)-L(\theta^*,w^*)\Big]\Big|\mathcal{F}_k\Big\}\\
&+\mathsf{E}\Big\{\Big[\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^*\Big\rangle-\Big\langle L'_\theta(\theta^{(k-1)},w^{(k)}),\,\theta^{(k)}\Big\rangle\Big]\Big|\mathcal{F}_k\Big\}
\end{aligned}
\tag{96}
$$

We will now bound the L terms in (96). First consider the second L term in (96):

$$
\begin{aligned}
&(n_X-1)\Big[L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\Big]\\
&\overset{(a)}{\leq}(n_X-1)\Big[\big\langle L'(\theta^{(k-1)},w^{(k)}-w^{(k-1)})\,,\theta^{(k-1)}-\theta^*\big\rangle\Big]\\
&\overset{(b)}{\leq}\beta_2(n_X-1)^2\|L'(\theta^{(k-1)},w^{(k)}-w^{(k-1)})\|^2+\frac{1}{\beta_2}\|\theta^{(k-1)}-\theta^*\|^2
\end{aligned}
\tag{97}
$$

for some $\beta_2>0$ to be determined later, where step (a) uses convexity of the function L in its first variable, and (b) uses $a^2+b^2\geq 2ab\geq ab$.

We further lower bound the first term in (97). Ignoring the scaling factors, this can be rewritten as follows:

$$
\begin{aligned}
\|L'(\theta^{(k-1)},w^{(k)}-w^{(k-1)})\|^2 &\overset{(a)}{=}\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta)(w_i^{(k)}-w_i^{(k-1)})\Big\|^2\\
&\overset{(b)}{=}\frac{1}{n_X^2}\Big\|\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta)(w_i^{(k)}-w_i^{(k-1)})\Big\|^2\\
&\overset{(c)}{\leq}\frac{B_f^2}{n_X^2}\Big\|\sum_{i=0}^{n_X-1}w_i^{(k)}-w_i^{(k-1)}\Big\|^2
\end{aligned}
\tag{98}
$$

where (a) follows from the definition (25), (b) is direct, by removing the $1/n_X$ outside the $\|\cdot\|^2$ operator, and (c) uses the bounded gradients assumption (Assumption E.3).

Notice that, conditioned on $\mathcal{F}_k$, $w^{(k)}$ is the only random variable in (98). Furthermore, for each i and j, using (38) we have,

$$
\sum_{i=0}^{n_X-1}w_i^{(k)}-w_i^{(k-1)}=w'_{ij}-w_i^{(k-1)}\qquad\text{with probability}\quad 1/n_X n_{Y_i}
\tag{99}
$$

Taking expectation, conditioned on $\mathcal{F}_k$ on both sides of (98),

$$\begin{aligned}
\mathsf{E}\Big\{\|L'(\theta^{(k-1)}, w^{(k)} - w^{(k-1)})\|^2 \Big| \mathcal{F}_k\Big\} \leq & \frac{B_f^2}{n_X^2}\mathsf{E}\Big\{\Big\| \sum_{i=0}^{n_X-1} w_i^{(k)} - w_i^{(k-1)}\Big\|^2 \Big| \mathcal{F}_k\Big\} \\
& \overset{(a)}{=} \frac{B_f^2}{n_X^2}\frac{1}{n_X n_{Y_i}} \sum_{i=0}^{n_X-1} \sum_{j=0}^{n_{Y_i}-1} \big\|w'_{ij} - w_i^{(k-1)}\big\|^2 \\
& \overset{(b)}{=} \frac{B_f^2}{n_X^2} \sum_{i=0}^{n_X-1} \mathsf{E}\Big\{\big\|w_i^{(k)} - w_i^{(k-1)}\big\|^2 \Big| \mathcal{F}_k\Big\} \\
& \overset{(c)}{=} \frac{B_f^2}{n_X^2}\mathsf{E}\Big\{\big\|w^{(k)} - w^{(k-1)}\big\|^2 \Big| \mathcal{F}_k\Big\}
\end{aligned} \tag{100}$$

where (a) follows from (99), (b) follows from the last identity in (52), and (c) is just definition. It follows from (97) and (100) that:

$$\begin{aligned}
& (n_X - 1)\mathsf{E}\Big\{\Big[L(\theta^{(k-1)}, w^{(k)} - w^{(k-1)}) - L(\theta^*, w^{(k)} - w^{(k-1)})\Big]\Big| \mathcal{F}_k\Big\} \\
\leq & \frac{\beta_2 B_f^2 (n_X - 1)^2}{n_X^2}\mathsf{E}\Big\{\big\|w^{(k)} - w^{(k-1)}\big\|^2 \Big| \mathcal{F}_k\Big\} + \frac{1}{\beta_2}\|\theta^{(k-1)} - \theta^*\|^2
\end{aligned} \tag{101}$$

Next consider the remaining L terms (lines 4 and 5 of (96)). We have:

$$\begin{aligned}
& L\big(\theta^{(k-1)},\, w^{(k)} - w^*\big) - L\big(\theta^*,\, w^{(k)} - w^*\big) + L(\theta^{(k)}, w^*) - L(\theta^*, w^*) \\
& = L\big(\theta^{(k-1)},\, w^{(k)}\big) - L\big(\theta^*,\, w^{(k)}\big) + L\big(\theta^*,\, w^*\big) - L\big(\theta^{(k-1)},\, w^*\big) + L(\theta^{(k)}, w^*) - L(\theta^*, w^*) \\
& = L\big(\theta^{(k-1)},\, w^{(k)}\big) - L\big(\theta^*,\, w^{(k)}\big) + L(\theta^{(k)}, w^*) - L\big(\theta^{(k-1)},\, w^*\big)
\end{aligned} \tag{102}$$

Therefore, the remaining L terms of (96) can be bounded as:

$$\begin{aligned}
& L(\theta^{(k)}, w^*) - L(\theta^*, w^*) + L\big(\theta^{(k-1)}, w^{(k)} - w^*\big) - L\big(\theta^*, w^{(k)} - w^*\big) - \langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^{(k)} - \theta^*\rangle \\
& \overset{(a)}{=} L\big(\theta^{(k-1)},\, w^{(k)}\big) - L\big(\theta^*,\, w^{(k)}\big) + L(\theta^{(k)}, w^*) - L\big(\theta^{(k-1)},\, w^*\big) - \langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^{(k)} - \theta^*\rangle \\
& \overset{(b)}{\leq} \big\langle L'_\theta(\theta^{(k-1)},\, w^{(k)}), \theta^{(k-1)} - \theta^*\big\rangle + \big\langle L'_\theta(\theta^{(k)}, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle - \langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^{(k)} - \theta^*\rangle \\
& \overset{(c)}{=} -\big\langle L'_\theta(\theta^{(k-1)},\, w^{(k)}), \theta^{(k)} - \theta^{(k-1)}\big\rangle + \big\langle L'_\theta(\theta^{(k)}, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \overset{(d)}{=} -\big\langle L'_\theta(\theta^{(k-1)},\, w^{(k)}), \theta^{(k)} - \theta^{(k-1)}\big\rangle + \big\langle L'_\theta(\theta^{(k-1)},\, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \quad - \big\langle L'_\theta(\theta^{(k-1)},\, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle + \big\langle L'_\theta(\theta^{(k)}, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \overset{(e)}{=} \big\langle L'_\theta(\theta^{(k-1)},\, w^*) - L'_\theta(\theta^{(k-1)},\, w^{(k)}), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \qquad + \big\langle L'_\theta(\theta^{(k)}, w^*) - L'_\theta(\theta^{(k-1)},\, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \overset{(f)}{=} -\big\langle L'_\theta(\theta^{(k-1)},\, w^{(k)} - w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle \\
& \qquad + \big\langle L'_\theta(\theta^{(k)}, w^*) - L'_\theta(\theta^{(k-1)},\, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle
\end{aligned} \tag{103}$$

where step (a) substitutes (102), step (b) uses the convexity of $L(\theta, w)$ in θ by applying $f(x) - f(y) \leq \langle f'(x), x - y\rangle$, step (c) merges the first and the third terms in line (b), step (d) adds and subtracts the same term (i.e., the second and the third terms), step (e) merges the first term with the second term and also merges the third term and the fourth term, step (f) uses the linearity of $L(\theta, w)$ in w. We now proceed to bound the two terms in (103). For a $\beta_1 > 0$ (to be determined later), the first term in (103) can be upper bounded as

$$\begin{aligned}
& \big|\langle L'_\theta(\theta^{(k-1)}, w^{(k)} - w^*),\, \theta^{(k)} - \theta^{(k-1)}\rangle\big| \\
& \overset{(a)}{\leq} \frac{1}{\beta_1}\big\|L'_\theta(\theta^{(k-1)}, w^{(k)} - w^*)\big\|^2 + \beta_1\big\|\theta^{(k)} - \theta^{(k-1)}\big\|^2
\end{aligned}$$

$$\overset{(b)}{\le} \frac{B_f^2}{\beta_1 n_X}\big\|w^{(k)} - w^*\big\|^2 + \beta_1\big\|\theta^{(k)} - \theta^{(k-1)}\big\|^2 \tag{104}$$

where (a) uses Cauchy-Schwartz inequality and the fact that $a^2 + b^2 \geq 2ab \geq ab$, and step (b) uses the definition of L'_θ in (25) and Jensen's inequality for $\|\cdot\|^2$. Next, the second term in (103) can be upper bounded as

$$\begin{aligned}
&\big|\big\langle L'_\theta(\theta^{(k)}, w^*) - L'_\theta(\theta^{(k-1)},\, w^*), \theta^{(k)} - \theta^{(k-1)}\big\rangle\big| \\
&\overset{(a)}{\le} \|L'_\theta(\theta^{(k)}, w^*) - L'_\theta(\theta^{(k-1)},\, w^*)\| \cdot \big\|\theta^{(k)} - \theta^{(k-1)}\big\| \\
&\overset{(b)}{\le} B_\theta B_w\|\theta^{(k)} - \theta^{(k-1)}\|^2
\end{aligned} \tag{105}$$

where step (a) uses Cauchy-Schwartz inequality and step (b) uses Lipschitz condition of the gradient f'_θ together with the boundedness of w^* (Lemma E.7). Substituting (104)–(105) into (103), we obtain

$$\begin{aligned}
&L(\theta^{(k)}, w^*) - L(\theta^*, w^*) + L\big(\theta^{(k-1)}, w^{(k)} - w^*\big) - L\big(\theta^*, w^{(k)} - w^*\big) - \langle L'_\theta(\theta^{(k-1)}, w^{(k)}), \theta^{(k)} - \theta^*\rangle \\
&\le \frac{B_f^2}{\beta_1 n_X}\big\|w^{(k)} - w^*\big\|^2 + \beta_1\big\|\theta^{(k)} - \theta^{(k-1)}\big\|^2 + B_\theta B_w\|\theta^{(k)} - \theta^{(k-1)}\|^2 \\
&= \frac{B_f^2}{\beta_1 n_X}\big\|w^{(k)} - w^*\big\|^2 + (\beta_1 + B_\theta B_w)\big\|\theta^{(k)} - \theta^{(k-1)}\big\|^2
\end{aligned} \tag{106}$$

We have now bounded all the L terms in (96).

Substituting both (101) and (106) in (96), we get the final bound, without the L terms as follows:

$$\begin{aligned}
&\left(\frac{1}{2\alpha_w} + \gamma\right)\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} + \frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 \mid \mathcal{F}_k\} \\
&\left(\frac{1}{2\alpha_\theta} + \mu\right)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2\Big|\mathcal{F}_k\Big\} + \frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)} - \theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\} \\
&\le \left(\frac{1}{2\alpha_w} + \frac{\gamma\big(n_X - 1\big)}{n_X}\right)\|w^{(k-1)} - w^*\|^2 + \frac{1}{2\alpha_\theta}\|\theta^{(k-1)} - \theta^*\|^2 \\
&+ \Big(2\alpha_w B_f^2\Big(1 - \overline{1/n_Y}\Big) + \alpha_\theta B_w^2 B_\theta^2\Big)\|\theta^{(k-1)} - \widetilde{\theta}\|^2 \\
&+ \frac{\beta_2 B_f^2(n_X - 1)^2}{n_X^2}\mathsf{E}\Big\{\|w^{(k)} - w^{(k-1)})\|^2\Big|\mathcal{F}_k\Big\} + \frac{1}{\beta_2}\|\theta^{(k-1)} - \theta^*\|^2 \\
&+ \frac{B_f^2}{\beta_1 n_X}\mathsf{E}\Big\{\big\|w^{(k)} - w^*\big\|^2\Big|\mathcal{F}_k\Big\} + (\beta_1 + B_\theta B_w)\mathsf{E}\Big\{\big\|\theta^{(k)} - \theta^{(k-1)}\big\|^2\Big|\mathcal{F}_k\Big\}
\end{aligned} \tag{107}$$

Combining common terms and rearranging,

$$\begin{aligned}
&\left(\frac{1}{2\alpha_\theta} + \mu\right)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2\Big|\mathcal{F}_k\Big\} + \left(\frac{1}{2\alpha_w} + \gamma - \frac{B_f^2}{\beta_1 n_X}\right)\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} \\
&\le \left(\frac{1}{2\alpha_w} + \frac{\gamma\big(n_X - 1\big)}{n_X}\right)\|w^{(k-1)} - w^*\|^2 \\
&+ \left(\frac{1}{2\alpha_\theta} + \frac{1}{\beta_2} + 4\alpha_w B_f^2\Big(1 - \overline{1/n_Y}\Big) + 2\alpha_\theta B_w^2 B_\theta^2\right)\|\theta^{(k-1)} - \theta^*\|^2 \\
&+ \left(\beta_1 + B_\theta B_w - \frac{1}{2\alpha_\theta}\right)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\} \\
&+ \left(\frac{\beta_2 B_f^2(n_X - 1)^2}{n_X^2} - \frac{1}{2\alpha_w}\right)\mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 \mid \mathcal{F}_k\} \\
&+ \Big(4\alpha_w B_f^2\Big(1 - \overline{1/n_Y}\Big) + 2\alpha_\theta B_w^2 B_\theta^2\Big)\|\widetilde{\theta} - \theta^*\|^2
\end{aligned} \tag{108}$$

where we have also used the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for the $\|\theta^{(k-1)} - \widetilde{\theta}\|^2$ term.

We now choose the step-sizes α_θ, α_w and M as in Theorem 4.6:

$$\alpha_\theta = \Big(\frac{25B_f^2}{\gamma} + 10B_\theta B_w + \frac{80B_w^2 B_\theta^2}{\mu}\Big)^{-1} \tag{109}$$

$$\alpha_w = \frac{\mu}{40B_f^2} \tag{110}$$

$$M = \max\left(\frac{10}{\alpha_\theta \mu}, \frac{2n_X}{\alpha_w \gamma}, 4n_X\right) \tag{111}$$

The choice of α_θ in (109) ensures the following three bounds:

$$\alpha_\theta \leq \frac{\gamma}{25B_f^2} \qquad \text{or} \qquad \alpha_\theta \leq \frac{1}{10B_\theta B_w} \qquad \text{or} \qquad \alpha_\theta \leq \frac{\mu}{80B_w^2 B_\theta^2} \tag{112}$$

Furthermore, we choose β_1 and β_2 as follows:

$$\beta_1 = -B_\theta B_w + \frac{1}{2\alpha_\theta} \tag{113}$$

$$\beta_2 = \frac{1}{2\alpha_w B_f^2} \tag{114}$$

The second inequality in (112) will ensure positivity of β_1.

Applying the above choice of hyper-parameters: Based on (113) we have:

$$\beta_1 + B_\theta B_w - \frac{1}{2\alpha_\theta} = 0 \tag{115}$$

and using (114):

$$\begin{aligned}\frac{\beta_2 B_f^2 (n_X - 1)^2}{n_X^2} - \frac{1}{2\alpha_w} &= \frac{1}{2\alpha_w B_f^2} \frac{B_f^2 (n_X-1)^2}{n_X^2} - \frac{1}{2\alpha_w} \\ &= \frac{1}{2\alpha_w} \frac{(n_X-1)^2}{n_X^2} - \frac{1}{2\alpha_w} \\ &< 0\end{aligned} \tag{116}$$

Equations (115) and (116) will ensure that the third and fourth terms on the right hand side of (108) are either 0 or negative, and therefore can be ignored, reducing the bound in (108) to:

$$\begin{aligned}&\left(\frac{1}{2\alpha_\theta} + \mu\right) \mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \big| \mathcal{F}_k\Big\} + \left(\frac{1}{2\alpha_w} + \gamma - \frac{B_f^2}{\beta_1 n_X}\right) \mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} \\ &\leq \left(\frac{1}{2\alpha_w} + \frac{\gamma\big(n_X - 1\big)}{n_X}\right) \|w^{(k-1)} - w^*\|^2 \\ &+ \left(\frac{1}{\beta_2} + \frac{1}{2\alpha_\theta} + 4\alpha_w B_f^2 + 2\alpha_\theta B_w^2 B_\theta^2\right) \|\theta^{(k-1)} - \theta^*\|^2 \\ &+ \Big(4\alpha_w B_f^2 + 2\alpha_\theta B_w^2 B_\theta^2\Big) \|\widetilde{\theta} - \theta^*\|^2\end{aligned} \tag{117}$$

where we have also used $(1 - \overline{1/n_Y}) \leq 1$.

The $\widetilde{\theta}$ in the (117) is $\widetilde{\theta}_{s-1}$, the fixed primal variable at the beginning of stage s. Denote $\widetilde{\theta}_s$ to be the primal variable randomly chosen among $\theta^{(k)}$ for $1 \leq k \leq M$ at the end of stage s. We define $\widetilde{w}_{s-1}$ and $\widetilde{w}_s$ in a similar manner (though neither of them are used in the algorithm), and also note that $\theta^{(0)}$

and $w^{(0)}$ are initialized to $\widetilde{\theta}_{s-1}$ and $\widetilde{w}_{s-1}$ at the beginning of stage s. Summing both sides of (117) over all $1 \leq k \leq M$,

$$\begin{aligned}
&\left(\frac{1}{2\alpha_\theta}+\mu\right)\mathsf{E}\Big\{\|\theta^{(M)}-\theta^*\|^2\big|\mathcal{F}_s\Big\}+\left(\frac{1}{2\alpha_w}+\gamma-\frac{B_f^2}{\beta_1 n_X}\right)\mathsf{E}\{\|w^{(M)}-w^*\|^2\mid\mathcal{F}_s\}\\
&+\left(\mu-\frac{1}{\beta_2}-4\alpha_w B_f^2-2\alpha_\theta B_w^2 B_\theta^2\right)\sum_{k=1}^{M-1}\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\big|\mathcal{F}_s\Big\}\\
&+\left(\frac{\gamma}{n_X}-\frac{B_f^2}{\beta_1 n_X}\right)\sum_{k=1}^{M-1}\mathsf{E}\{\|w^{(k)}-w^*\|^2\mid\mathcal{F}_s\}\\
&\leq\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(0)}-w^*\|^2+\left(\frac{1}{\beta_2}+\frac{1}{2\alpha_\theta}\right)\|\theta^{(0)}-\theta^*\|^2\\
&+M\Big(4\alpha_w B_f^2+2\alpha_\theta B_w^2 B_\theta^2\Big)\|\widetilde{\theta}-\theta^*\|^2
\end{aligned}\tag{118}$$

Substituting $w^{(0)}=\widetilde{w}_{s-1}$ and $\theta^{(0)}=\widetilde{\theta}_{s-1}$, and also noting that the first two terms on the left hand side of (118) can be combined with the second two terms (note that the difference in coefficients are positive, and positive terms on the left hand side of the inequality can be ignored):

$$\begin{aligned}
&\left(\mu-\frac{1}{\beta_2}-4\alpha_w B_f^2-2\alpha_\theta B_w^2 B_\theta^2\right)\sum_{k=1}^{M}\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\big|\mathcal{F}_s\Big\}\\
&+\left(\frac{\gamma}{n_X}-\frac{B_f^2}{\beta_1 n_X}\right)\sum_{k=1}^{M}\mathsf{E}\{\|w^{(k)}-w^*\|^2\mid\mathcal{F}_s\}\\
&\leq\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|\widetilde{w}_{s-1}-w^*\|^2\\
&+\left(\frac{1}{\beta_2}+\frac{1}{2\alpha_\theta}+M\Big(4\alpha_w B_f^2+2\alpha_\theta B_w^2 B_\theta^2\Big)\right)\|\widetilde{\theta}_{s-1}-\theta^*\|^2
\end{aligned}\tag{119}$$

Dividing both sides of (119) by M and applying Jensen's inequality on the left hand side, we obtain:

$$\begin{aligned}
&\left(\mu-\frac{1}{\beta_2}-4\alpha_w B_f^2-2\alpha_\theta B_w^2 B_\theta^2\right)\mathsf{E}\Big\{\|\widetilde{\theta}_s-\theta^*\|^2\big|\mathcal{F}_s\Big\}\\
&+\left(\frac{\gamma}{n_X}-\frac{B_f^2}{\beta_1 n_X}\right)\mathsf{E}\{\|\widetilde{w}_s-w^*\|^2\mid\mathcal{F}_s\}\\
&\leq\left(\frac{1}{2M\alpha_w}+\frac{\gamma\big(n_X-1\big)}{Mn_X}\right)\|\widetilde{w}_{s-1}-w^*\|^2\\
&+\left(\frac{1}{M\beta_2}+\frac{1}{2M\alpha_\theta}+4\alpha_w B_f^2+2\alpha_\theta B_w^2 B_\theta^2\Big)\|\widetilde{\theta}_{s-1}-\theta^*\|^2\right.
\end{aligned}\tag{120}$$

We now substitute the hyper-parameter values in (109) (110) (111) (113) (114) in (120) to obtain a linear rate.

Substituting these values in the coefficient of the first term on the left hand side of (120) (we're using the third bound for α_θ in (112) here):

$$\begin{aligned}
\mu - \frac{1}{\beta_2} - 4\alpha_w B_f^2 - 2\alpha_\theta B_w^2 B_\theta^2 &= \mu - 2\alpha_w B_f^2 - 4\alpha_w B_f^2 - 2\alpha_\theta B_w^2 B_\theta^2 \\
&\geq \mu - 2\frac{\mu}{40B_f^2}B_f^2 - 4\frac{\mu}{40B_f^2}B_f^2 - 2\frac{\mu}{80B_w^2 B_\theta^2}B_w^2 B_\theta^2 \\
&= \mu - \frac{\mu}{20} - \frac{\mu}{10} - \frac{\mu}{40} \\
&= \mu - \frac{7\mu}{40} \\
&= \frac{33\mu}{40} \\
&\geq \frac{4\mu}{5}
\end{aligned} \tag{121}$$

Next, substituting for the coefficient of the second term on the left hand side of (120) (here we use the first and second bounds for α_θ in (112), both in the inequality in the fourth line):

$$\begin{aligned}
\frac{\gamma}{n_X} - \frac{B_f^2}{\beta_1 n_X} &= \frac{\gamma}{n_X} - \frac{B_f^2}{n_X}\left(- B_\theta B_w + \frac{1}{2\alpha_\theta}\right)^{-1} \\
&= \frac{\gamma}{n_X} - \frac{B_f^2}{n_X}\left(\frac{1 - 2\alpha_\theta B_\theta B_w}{2\alpha_\theta}\right)^{-1} \\
&= \frac{\gamma}{n_X} - \frac{B_f^2}{n_X}\left(\frac{2\alpha_\theta}{1 - 2\alpha_\theta B_\theta B_w}\right) \\
&\geq \frac{\gamma}{n_X} - \frac{B_f^2}{n_X} \times \frac{2\gamma}{25B_f^2} \times \frac{1}{1 - \frac{2}{10B_\theta B_w}B_\theta B_w} \\
&= \frac{\gamma}{n_X} - \frac{2\gamma}{25n_X} \times \frac{5}{4} \\
&= \frac{\gamma}{n_X} - \frac{\gamma}{10n_X} \\
&\geq \frac{4\gamma}{5n_X}
\end{aligned} \tag{122}$$

Next consider the coefficient of the first term on the right hand side of (120) (we use the second and third values of M in (111), both in the second line):

$$\begin{aligned}
\frac{1}{2M\alpha_w} + \frac{\gamma\big(n_X - 1\big)}{Mn_X} &\leq \frac{1}{2M\alpha_w} + \frac{\gamma}{M} \\
&\leq \frac{\gamma}{4n_X} + \frac{\gamma}{4n_X} \\
&= \frac{\gamma}{2n_X}
\end{aligned} \tag{123}$$

Finally, we consider the coefficient of the second term on the right hand side of (120) (we use the third bound for α_θ in (112), and the first and third definitions of M in (111)):

$$\begin{aligned}
\frac{1}{M\beta_2} + \frac{1}{2M\alpha_\theta} + 4\alpha_w B_f^2 + 2\alpha_\theta B_w^2 B_\theta^2 &\leq \frac{2\alpha_w B_f^2}{M} + \frac{\mu}{20} + 4\frac{\mu}{40B_f^2}B_f^2 + \frac{\mu}{80B_w^2 B_\theta^2}2B_w^2 B_\theta^2 \\
&\leq \frac{\mu}{40B_f^2}2B_f^2 + \frac{\mu}{20} + \frac{\mu}{10} + \frac{\mu}{40} \\
&= \frac{9\mu}{40} \\
&= \frac{\mu}{4}
\end{aligned} \tag{124}$$

Using (121), (122), (123), and (124) in (120), we have:

$$\begin{aligned}
&\frac{4\mu}{5}\mathsf{E}\Big\{\|\widetilde{\theta}_s-\theta^*\|^2\big|\mathcal{F}_s\Big\}+\frac{4\gamma}{5n_X}\mathsf{E}\{\|\widetilde{w}_s-w^*\|^2\mid\mathcal{F}_s\}\\
&\le\frac{\gamma}{2n_X}\|\widetilde{w}_{s-1}-w^*\|^2+\frac{\mu}{4}\|\widetilde{\theta}_{s-1}-\theta^*\|^2
\end{aligned}\tag{125}$$

Applying full expectation to both sides of the above inequality, we obtain

$$\begin{aligned}
&\frac{4\mu}{5}\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2+\frac{4\gamma}{5n_X}\mathsf{E}\|\widetilde{w}_s-w^*\|^2\\
&\le\frac{\gamma}{2n_X}\mathsf{E}\|\widetilde{w}_{s-1}-w^*\|^2+\frac{\mu}{4}\mathsf{E}\|\widetilde{\theta}_{s-1}-\theta^*\|^2.
\end{aligned}\tag{126}$$

Dividing both sides by $4\mu/5$, we have

$$\begin{aligned}
&\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2+\frac{\gamma}{n_X\mu}\mathsf{E}\|\widetilde{w}_s-w^*\|^2\\
&\le\frac{5}{16}\mathsf{E}\|\widetilde{w}_{s-1}-w^*\|^2+\frac{5\gamma}{8\mu n_X}\mathsf{E}\|\widetilde{\theta}_{s-1}-\theta^*\|^2,
\end{aligned}\tag{127}$$

which can be further bounded as

$$\begin{aligned}
&\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2+\frac{\gamma}{n_X\mu}\mathsf{E}\|\widetilde{w}_s-w^*\|^2\\
&\le\frac{5}{8}\Big(\mathsf{E}\|\widetilde{w}_{s-1}-w^*\|^2+\frac{\gamma}{\mu n_X}\mathsf{E}\|\widetilde{\theta}_{s-1}-\theta^*\|^2\Big).
\end{aligned}\tag{128}$$

Define the Lyapunov function P_s to be

$$P_s=\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2+\frac{\gamma}{n_X\mu}\mathsf{E}\|\widetilde{w}_s-w^*\|^2.$$

Then, inequality (126) can be expressed as

$$P_s\le\frac{5}{8}P_{s-1}\le\Big(\frac{5}{8}\Big)^sP_0.\tag{129}$$

Therefore, P_s converges to zero at a linear rate of $5/8$. In order to achieve ϵ-precision solution (i.e., $P_s\le\epsilon$), it requires a total of $O(\ln\frac{1}{\epsilon})$ outer-loop iterations (stages). And for each outer loop iteration, it requires M steps of inner-loop primal-dual updates, which is $O(1)$ per step (in number of oracle calls), and $O(n_Xn_Y)$ for evaluating the batch gradients for the control variates, where $n_Y=(n_{Y_0}+\cdots+n_{Y_{n_X-1}})/n_X$. Therefore, the complexity per outer loop iteration is $O(n_Xn_Y+M)$ so that the total complexity can be written as:

$$O\Big(\big(n_Xn_Y+M\big)\ln\big(\frac{P_0}{\epsilon}\big)\Big).\tag{130}$$

Recall from (111) that M is given by

$$M=10/\mu\alpha_\theta+\frac{2n_X}{\alpha_w\gamma}+4n_X.$$

where, by (109) and (110), the step-sizes α_θ and θ_w are given by

$$\alpha_\theta=\Big(\frac{25B_f^2}{\gamma}+10B_\theta B_w+\frac{80B_w^2B_\theta^2}{\mu}\Big)^{-1},\quad\alpha_w=\frac{\mu}{40B_f^2}.$$

This implies that $M=O(B_f^2/\mu\gamma+B_w^2B_\theta^2/\mu^2+(B_f^2/\mu\gamma)n_X+n_X)$. In consequence, the total complexity is

$$O\Big(\big(n_Xn_Y+n_X\kappa+n_X\big)\ln\frac{1}{\epsilon}\Big),\tag{131}$$

where

$$\kappa=B_f^2/\gamma\mu+B_w^2B_\theta^2/\mu^2.\tag{132}$$

Noting that $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2\le P_s$, the bound (129) implies that $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2$ also converges to zero at a linear rate of $5/8$ and the total complexity to reach $\mathsf{E}\|\widetilde{\theta}_s-\theta^*\|^2\le\epsilon$ is also given by (131).

F Special case: SVRPDA-I with $n_{Y_i} \equiv 1$, f_θ linear in θ and no outer loop

First, we observe that in this special case, our SVRPDA-I algorithm will become a single-loop algorithm, and that the outer-loop in Algorithm 1 is no longer needed. To see this, first note that when $n_{Y_i} \equiv 1$, δ_k^w is independent of $\widetilde{\theta}$ because the last two terms in (12) would cancel each other. Second, when f_θ is linear in θ, the term $\overline{f}'_{i_k}(\widetilde{\theta})$ in (14) and U_0 in (11) are independent of $\widetilde{\theta}$, which further implies that U_k (that is recursively defined in (14)) is also independent of $\widetilde{\theta}$. Finally, we also note that, with linear f_θ, the first two terms in (15) cancel with each other, so that $\delta_k^\theta \equiv U_k$ is independent of $\widetilde{\theta}$. As a result, the inner loop in Algorithm 1 does not require an outer-loop to update the reference variable $\widetilde{\theta}$.

The following theorem establishes the complexity bound for the SVRPDA-I algorithm in this special case.

Theorem F.1. *Suppose Assumptions 4.1–4.4 hold. Furthermore, suppose $n_{Y_i} = 1$, $1 \le i \le n_X$, and f_θ is a linear function of θ. Consider just the the inner loop of Algorithm 1, with s = 1 fixed, and*

$$\alpha_\theta = \frac{\gamma}{16B_f^2 + 4n_X\mu\gamma}, \quad \textit{and} \quad \alpha_w = \frac{1}{2\gamma}\frac{3n_X + \kappa + 1}{n_X + \kappa + 1}$$

where $\kappa = B_f^2/\gamma\mu$ is the condition number. Then, the Lyapunov function

$$\Delta^{(k)} := \Big(\frac{1}{2\alpha_\theta} + \mu\Big)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2\Big|\mathcal{F}_k\Big\} + \Big(\frac{1}{2\alpha_w} + \gamma\Big)\mathsf{E}\Big\{\|w^{(k)} - w^*\|^2\Big|\mathcal{F}_k\Big\}$$

satisfies $\Delta^{(k)} \le \big(1 - 1/\big(1 + 2\kappa + 2n_X\big)\big)^k\Delta^{(0)}$. Furthermore, the overall computational cost (in number of oracle calls) for reaching $\Delta^{(k)} \le \epsilon$ is upper bounded by

$$O\Big(\big(n_X + \kappa\big)\ln\Big(\frac{1}{\epsilon}\Big)\Big). \tag{133}$$

In comparison, the authors in [5] propose a stochastic primal dual coordinate (SPDC) algorithm for this special case and prove an overall complexity of $O\big(\big(n_X + \sqrt{n_X\kappa}\big)\ln\big(\frac{1}{\epsilon}\big)\big)$ to achieve an ϵ-error solution, where the condition number $\kappa = B_f^2/\mu\gamma$. This is by far the best complexity for this class of problems. It is interesting to note that the complexity result in (133) and the complexity result in [5] only differ in their dependency on κ. This difference is most likely due to the acceleration technique that is employed in the primal update of the SPDC algorithm. We conjecture that the dependency on the condition number of SVRPDA-I can be further improved using a similar acceleration technique.

F.1 Proof of Theorem F.1

It is useful to first discuss the main implications of choosing f_θ to be linear in θ and $n_{Y_i} = 1$ for all i. First, based on Assumption E.3 (or equivalently **Assumption 4.3**), we have $B_\theta = 0$, since f'_θ is independent of θ. This also implies that L'_θ is independent of θ, and therefore, Assumption E.4 (or equivalently **Assumption 4.4**) holds with equality:

$$L(\theta_1, w) - L(\theta_2, w) = \langle L'_\theta(\theta_2, w),\ \theta_1 - \theta_2\rangle. \tag{134}$$

In particular, for any $\theta \in \mathbb{R}^d$ and $w \in \mathbb{R}^\ell$, $L(\theta, w) = \langle L'_\theta(\theta, w),\ \theta\rangle$. Finally, $n_{Y_i} = 1$ implies $\overline{1/n_Y} = \frac{1}{n_X}\sum_{i=0}^{n_X-1} 1/n_{Y_i} = 1$.

Using the above implications in the primal bound (71) (in particular, letting $B_\theta = 0$, and using linearity of $L(\theta, w)$), we obtain the primal bound for the special case as follows:

$$\begin{aligned}&\left(\frac{1}{2\alpha_\theta} + \mu\right)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2\big|\mathcal{F}_k\Big\} + \frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)} - \theta^{(k-1)}\|^2\big|\mathcal{F}_k\Big\}\\ \le&\frac{1}{2\alpha_\theta}\|\theta^{(k-1)} - \theta^*\|^2 - \mathsf{E}\Big\{\Big[L(\theta^{(k)}, w^{(k)} - w^*) - L(\theta^*, w^{(k)} - w^*)\Big]\big|\mathcal{F}_k\Big\}\end{aligned} \tag{135}$$

Similarly, using the above implications in the dual bound (56) (in particular, letting $(1 - \overline{1/n_Y}) = 0$), the dual bound for the special case becomes:

$$\left(\frac{1}{2\alpha_w} + \gamma\right)\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} + \frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)} - w^{(k-1)}\|^2 \mid \mathcal{F}_k\}$$

$$\begin{aligned}
\le&\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(k-1)}-w^*\|^2\\
&+\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k-1)}-w^*)-L(\theta^*,w^{(k-1)}-w^*)\mid\mathcal{F}_k\Big\}\\
&+n_X\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\mid\mathcal{F}_k\Big\}
\end{aligned}\tag{136}$$

Adding (135) and (136), we obtain the combined primal-dual bound:

$$\begin{aligned}
&\left(\frac{1}{2\alpha_\theta}+\mu\right)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\Big|\mathcal{F}_k\Big\}+\frac{1}{2\alpha_\theta}\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\}\\
&+\left(\frac{1}{2\alpha_w}+\gamma\right)\mathsf{E}\{\|w^{(k)}-w^*\|^2\mid\mathcal{F}_k\}+\frac{1}{2\alpha_w}\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2\mid\mathcal{F}_k\}\\
\le&\frac{1}{2\alpha_\theta}\|\theta^{(k-1)}-\theta^*\|^2+\left(\frac{1}{2\alpha_w}+\frac{\gamma\big(n_X-1\big)}{n_X}\right)\|w^{(k-1)}-w^*\|^2\\
&-\mathsf{E}\Big\{\Big[L(\theta^{(k)},w^{(k)}-w^*)-L(\theta^*,w^{(k)}-w^*)\Big]\Big|\mathcal{F}_k\Big\}\\
&+\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k-1)}-w^*)-L(\theta^*,w^{(k-1)}-w^*)\mid\mathcal{F}_k\Big\}\\
&+n_X\mathsf{E}\Big\{L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\mid\mathcal{F}_k\Big\}
\end{aligned}\tag{137}$$

As done in the previous proofs, we will first consider all the L terms that appear on the right hand side of (137). Following exactly the same steps as (73), (74), (75) and (76), we obtain the final bound for the L terms as given in (77), but with $B_\theta=0$. We still write out the whole simplification details here for completeness, and also to show how this special case is much simpler than the more general case.

Considering all the L terms in (137), we have:

$$\begin{aligned}
&-\Big[L(\theta^{(k)},w^{(k)}-w^*)-L(\theta^*,w^{(k)}-w^*)\Big]+\Big[L(\theta^{(k-1)},w^{(k-1)}-w^*)-L(\theta^*,w^{(k-1)}-w^*)\Big]\\
&+n_X\Big[L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\Big]\\
=&\Big[L(\theta^{(k-1)},w^{(k-1)}-w^*)-L(\theta^{(k)},w^{(k-1)}-w^*)\Big]\\
&+\Big[L(\theta^{(k)},w^{(k-1)}-w^{(k)})-L(\theta^*,w^{(k-1)}-w^{(k)})\Big]\\
&+n_X\Big[L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\Big]\\
=&\Big[L(\theta^{(k-1)},w^{(k-1)}-w^*)-L(\theta^{(k)},w^{(k-1)}-w^*)\Big]\\
&+\Big[L(\theta^{(k)},w^{(k-1)}-w^{(k)})-L(\theta^{(k-1)},w^{(k-1)}-w^{(k)})\Big]\\
&+(n_X-1)\Big[L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\Big]\\
=&\Big[L(\theta^{(k-1)},w^{(k)}-w^*)-L(\theta^{(k)},w^{(k)}-w^*)\Big]\\
&+(n_X-1)\Big[L(\theta^{(k-1)},w^{(k)}-w^{(k-1)})-L(\theta^*,w^{(k)}-w^{(k-1)})\Big]\\
=&L\Big(\theta^{(k-1)},w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big)-L\Big(\theta^{(k)},w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big)\\
\overset{(a)}{\le}&\Big\langle L'_\theta\Big(\theta^{(k-1)},w^{(k)}-w^*+(n_X-1)(w^{(k)}-w^{(k-1)})\Big),\theta^{(k-1)}-\theta^{(k)}\Big\rangle\\
=&\Big\langle\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k)}-w_i^*+(n_X-1)(w_i^{(k)}-w_i^{(k-1)})\Big),\theta^{(k-1)}-\theta^{(k)}\Big\rangle\\
=&\Big\langle\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}+n_X(w_i^{(k)}-w_i^{(k-1)})-w_i^*\Big),\theta^{(k-1)}-\theta^{(k)}\Big\rangle
\end{aligned}$$

$$\begin{aligned}
&\overset{(b)}{\le}\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}+n_X(w_i^{(k)}-w_i^{(k-1)})-w_i^*\Big)\Big\|^2+\frac{1}{\beta_1}\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2\\
&=\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\big(w_i^{(k-1)}-w_i^*\big)+\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2+\frac{1}{\beta_1}\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2\\
&\overset{(c)}{\le}2\beta_1\Big\|\frac{1}{n_X}\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})\Big(w_i^{(k-1)}-w_i^*\Big)\Big\|^2+2\beta_1\Big\|\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2\\
&\quad+\frac{1}{\beta_1}\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2\\
&\overset{(d)}{\le}\frac{2\beta_1B_f^2}{n_X}\sum_{i=0}^{n_X-1}\|w_i^{(k-1)}-w_i^*\|^2+2\beta_1\Big\|\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2+\frac{1}{\beta_1}\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2
\end{aligned}\tag{138}$$

where step (a) uses convexity of L in θ, step (b) uses $\langle a,b\rangle\le\beta_1\|a\|^2+\frac{1}{\beta_1}\|b\|^2$, step (c) uses $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, and step (d) applies Jensen's inequality and bounded gradient assumption to the first term. The second term in (138) can be simplified as follows:

$$\begin{aligned}
\mathsf{E}\Big\{\Big\|\sum_{i=0}^{n_X-1}\overline{f}'_i(\theta^{(k-1)})(w_i^{(k)}-w_i^{(k-1)})\Big\|^2\Big|\mathcal{F}_k\Big\}&=\sum_{i=0}^{n_X-1}\sum_{j=0}^{n_{Y_i}-1}\frac{1}{n_Xn_{Y_i}}\Big\|\overline{f}'_i(\theta^{(k-1)})(w'_{ij}-w_i^{(k-1)})\Big\|^2\\
&\le B_f^2\sum_{i=0}^{n_X-1}\sum_{j=0}^{n_{Y_i}-1}\frac{1}{n_Xn_{Y_i}}\Big\|w'_{ij}-w_i^{(k-1)}\Big\|^2\\
&=B_f^2\sum_{i=0}^{n_X-1}\mathsf{E}\{\|w_i^{(k)}-w_i^{(k-1)}\|^2\}
\end{aligned}\tag{139}$$

Therefore, using (138) and (139), the conditional expectation of all the L terms in (137) are bounded by:

$$\begin{aligned}
&\frac{2\beta_1B_f^2}{n_X}\sum_{i=0}^{n_X-1}\|w_i^{(k-1)}-w_i^*\|^2+2\beta_1B_f^2\sum_{i=0}^{n_X-1}\mathsf{E}\{\|w_i^{(k)}-w_i^{(k-1)}\|^2|\mathcal{F}_k\}+\frac{1}{\beta_1}\mathsf{E}\{\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2|\mathcal{F}_k\}\\
&=\frac{2\beta_1B_f^2}{n_X}\|w^{(k-1)}-w^*\|^2+2\beta_1B_f^2\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2|\mathcal{F}_k\}+\frac{1}{\beta_1}\mathsf{E}\{\big\|\theta^{(k-1)}-\theta^{(k)}\big\|^2|\mathcal{F}_k\}
\end{aligned}\tag{140}$$

Substituting the above upper bound for the L terms in (137) leads to

$$\begin{aligned}
&\left(\frac{1}{2\alpha_\theta}+\mu\right)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^*\|^2\Big|\mathcal{F}_k\Big\}+\left(\frac{1}{2\alpha_w}+\gamma\right)\mathsf{E}\{\|w^{(k)}-w^*\|^2\mid\mathcal{F}_k\}\\
&\quad+\Big(\frac{1}{2\alpha_\theta}-\frac{1}{\beta_1}\Big)\mathsf{E}\Big\{\|\theta^{(k)}-\theta^{(k-1)}\|^2\Big|\mathcal{F}_k\Big\}+\Big(\frac{1}{2\alpha_w}-2\beta_1B_f^2\Big)\mathsf{E}\{\|w^{(k)}-w^{(k-1)}\|^2\mid\mathcal{F}_k\}\\
&\le\frac{1}{2\alpha_\theta}\|\theta^{(k-1)}-\theta^*\|^2+\left(\frac{1}{2\alpha_w}+\gamma-\frac{\gamma}{n_X}+\frac{2\beta_1B_f^2}{n_X}\right)\|w^{(k-1)}-w^*\|^2
\end{aligned}\tag{141}$$

which will be the final bound we will analyze.

Substituting hyper-parameter choices

Recall the choice of step-sizes α_θ and α_w defined in Theorem F.1:

$$\begin{aligned}
\alpha_\theta&=\frac{\gamma}{16B_f^2+4n_X\mu\gamma}\\
\alpha_w&=\frac{1}{2\gamma}\frac{3n_X+\kappa+1}{n_X+\kappa+1}
\end{aligned}\tag{142}$$

where the condition number κ is also defined in Theorem F.1:

$$\kappa = B_f^2/\gamma\mu$$

Furthermore, choosing β_1 such that

$$\frac{1}{2\alpha_\theta} = \frac{1}{\beta_1}, \tag{143}$$

we have:

$$\begin{aligned}
\alpha_\theta\alpha_w &= \frac{\gamma}{16B_f^2 + 4n_X\mu\gamma} \times \frac{1}{2\gamma}\frac{3n_X+\kappa+1}{n_X+\kappa+1} \\
&= \frac{1}{32B_f^2 + 8n_X\mu\gamma} \times \frac{3n_X+\kappa+1}{n_X+\kappa+1} \\
&< \frac{1}{32B_f^2 + 8n_X\mu\gamma} \times \frac{3n_X+3\kappa+3}{n_X+\kappa+1} \\
&= \frac{3}{32B_f^2 + 8n_X\mu\gamma} \\
&< \frac{1}{8B_f^2 + 2n_X\mu\gamma} \\
&< \frac{1}{8B_f^2}
\end{aligned} \tag{144}$$

so that,

$$\begin{aligned}
&\frac{1}{2\alpha_w} > 4\alpha_\theta B_f^2 \\
\implies &\frac{1}{2\alpha_w} > 2\beta_1 B_f^2
\end{aligned} \tag{145}$$

Identities (143) and (145) make sure that the third and the fourth terms on the left hand side of (141) are either zero or positive, so that they can be ignored, resulting in:

$$\begin{aligned}
&\left(\frac{1}{2\alpha_\theta} + \mu\right)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2\Big|\mathcal{F}_k\Big\} + \left(\frac{1}{2\alpha_w} + \gamma\right)\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\} \\
\le&\frac{1}{2\alpha_\theta}\|\theta^{(k-1)} - \theta^*\|^2 + \left(\frac{1}{2\alpha_w} + \gamma - \frac{\gamma}{n_X} + \frac{2\beta_1 B_f^2}{n_X}\right)\|w^{(k-1)} - w^*\|^2
\end{aligned} \tag{146}$$

We further look at the coefficients of other error terms in (146). To this end, we have:

$$\begin{aligned}
\frac{1}{2\alpha_\theta\mu} &= \frac{16B_f^2 + 4n_X\mu\gamma}{2\mu\gamma} \\
&= \frac{8B_f^2}{\mu\gamma} + 2n_X \\
&= 2\kappa + 2n_X
\end{aligned} \tag{147}$$

and therefore, letting r_P to denote the ratio of the coefficients of $\|\theta^{(k-1)} - \theta^*\|^2$ and $\mathsf{E}\big\{\|\theta^{(k)} - \theta^*\|^2\big|\mathcal{F}_k\big\}$, we have:

$$\begin{aligned}
r_P &:= \Big(\frac{1}{2\alpha_\theta}\Big)\Big/\Big(\frac{1}{2\alpha_\theta} + \mu\Big) \\
&= 1 - \frac{2\alpha_\theta\mu}{1 + 2\alpha_\theta\mu}, \\
&= 1 - \frac{1}{1 + \frac{1}{2\alpha_\theta\mu}} \\
&= 1 - \frac{1}{1 + 2\kappa + 2n_X}
\end{aligned} \tag{148}$$

Also, for the dual terms we have r_D denoting the ratio of the coefficients of $\|w^{(k-1)} - w^*\|^2$ and $\mathsf{E}\{\|w^{(k)} - w^*\|^2 \mid \mathcal{F}_k\}$, so that:

$$\begin{aligned} r_D &:= \left(\frac{1}{2\alpha_w} + \gamma - \frac{\gamma}{n_X} + \frac{2\beta_1 B_f^2}{n_X}\right) \Big/ \left(\frac{1}{2\alpha_w} + \gamma\right) \\ &= 1 - \frac{1}{n_X}\frac{2\alpha_w\gamma - 8\alpha_w\alpha_\theta B_f^2}{1 + 2\alpha_w\gamma} \\ &\overset{(a)}{\leq} 1 - \frac{1}{n_X}\frac{2\alpha_w\gamma - 1}{2\alpha_w\gamma + 1} \\ &\overset{(b)}{=} 1 - \frac{1}{n_X}\frac{2n_X}{4n_X + 2\kappa + 2} \\ &= 1 - \frac{1}{2n_X + \kappa + 1} \end{aligned} \tag{149}$$

where step (a) follows from (144), and step (b) follows from (150) below:

$$\begin{aligned} \frac{2\alpha_w\gamma - 1}{2\alpha_w\gamma + 1} &= \frac{\frac{3n_X+\kappa+1}{n_X+\kappa+1} - 1}{\frac{3n_X+\kappa+1}{n_X+\kappa+1} + 1} \\ &= \frac{3n_X + \kappa + 1 - n_X - \kappa - 1}{3n_X + \kappa + 1 + n_X + \kappa + 1} \\ &= \frac{2n_X}{4n_X + 2\kappa + 2} \end{aligned} \tag{150}$$

From (148), the above bound (149) on r_D implies $r_D \leq r_P$.

Recalling the definition of $\Delta^{(k)}$, the Lyapunov function defined in Theorem F.1; Substituting for the step-size values in (142), we have:

$$\begin{aligned} \Delta^{(k)} &= \Big(\frac{1}{2\alpha_\theta} + \mu\Big)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \Big| \mathcal{F}_k\Big\} + \Big(\frac{1}{2\alpha_w} + \gamma\Big)\mathsf{E}\Big\{\|w^{(k)} - w^*\|^2 \Big| \mathcal{F}_k\Big\} \\ &= \Big(\mu\big(2\kappa + 2n_X + 1\big)\Big)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \Big| \mathcal{F}_k\Big\} + \Big(\gamma\Big(\frac{n_X + \kappa + 1}{3n_X + \kappa + 1} + 1\Big)\Big)\mathsf{E}\Big\{\|w^{(k)} - w^*\|^2 \Big| \mathcal{F}_k\Big\} \\ &= \Big(\mu\big(2\kappa + 2n_X + 1\big)\Big)\mathsf{E}\Big\{\|\theta^{(k)} - \theta^*\|^2 \Big| \mathcal{F}_k\Big\} + \Big(\gamma\Big(\frac{4n_X + 2\kappa + 2}{3n_X + \kappa + 1}\Big)\Big)\mathsf{E}\Big\{\|w^{(k)} - w^*\|^2 \Big| \mathcal{F}_k\Big\} \end{aligned}$$

Based on (148) and (149), and the fact that $r_D \leq r_P$, the inequality (141) then implies:

$$\Delta^{(k)} \leq r\Delta^{(k-1)} \tag{151}$$

where,

$$r = r_P = 1 - 1/\big(1 + 2\kappa + 2n_X\big). \tag{152}$$

The bound (151) implies that after k iterations, the error $\Delta^{(k)}$ satisfies:

$$\Delta^{(k)} \leq r^k \Delta^{(0)}$$

Therefore, for $\Delta^{(k)} < \epsilon$, it suffices to have $r^k\Delta^{(0)} < \epsilon$, so that the number of iterations k satisfies:

$$\begin{aligned} & k \ln r \leq \ln\Big(\frac{\epsilon}{\Delta^{(0)}}\Big) \\ \Longrightarrow\ & k \ln\Big(1 - 1/\big(1 + 2\kappa + 2n_X\big)\Big) \leq \ln\Big(\frac{\epsilon}{\Delta^{(0)}}\Big) \\ \Longrightarrow\ & -k/\big(1 + 2\kappa + 2n_X\big) \leq \ln\Big(\frac{\epsilon}{\Delta^{(0)}}\Big) \\ \Longrightarrow\ & k \geq \Big(1 + 2\kappa + 2n_X\Big)\ln\Big(\frac{\Delta^{(0)}}{\epsilon}\Big) \end{aligned} \tag{153}$$

where we have used $-\ln(1 - x) \geq x$. (153) implies the final complexity result of Theorem F.1. ☐

References

[1] X. Lian, M. Wang, and J. Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 1159–1167, 2017.

[2] M. Wang, J. Liu, and E. Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*, pages 1714–1722, 2016.

[3] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[4] J. Zhang and L. Xiao. A composite randomized incremental gradient method. In *International Conference on Machine Learning*, pages 7454–7462, 2019.

[5] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(1):2939–2980, 2017.