Stochastic Bilevel Optimization

February 20, 2023

1 Problem

We set up a bilevel problem. For a bounded domain $\mathcal{D} \subset \mathbb{R}^r$,

$$\min_{\lambda \in \mathcal{D}} \mathcal{L}(\lambda) \triangleq C(\hat{x}(\lambda), \lambda) \tag{1}$$

$$\text{s.t. } \hat{x}(\lambda) = \underset{x \in \mathbb{R}^n}{\arg\min} F(x, \lambda). \tag{2}$$

We will denote the sampled terms as follows. For i.i.d. samples ζ, ξ, uniformly randomly selected from the given data $\{\zeta_i, \xi_j, i = 1, \ldots, m_1; j = 1, \ldots, m_2\}$, we have:

$$\mathbb{E}_\zeta[\tilde{F}(x, \lambda; \zeta)] = F(x, \lambda)$$
$$\mathbb{E}_\xi[\tilde{C}(x, \lambda, \xi)] = C(x, \lambda)$$

For a function $f : \mathbb{R}^n \to \mathbb{R}$, will denote the conditional expectation $\mathbb{E}_{x_k}[f(x)] = \mathbb{E}[f(x)|x = x_k]$.

2 Algorithm

Algorithm 1 Stochastic HOAG

1: At iteration $k = 1, 2, \ldots$, choose batches of i.i.d. uniform random samples $\mathcal{S}_{\nabla_x C}, \mathcal{S}_{\nabla_\lambda C}, \mathcal{S}_{\nabla_{x\lambda} F}$, batches of i.i.d. random samples $\mathcal{B}_j$, stepsize ν_k, parameter η, perform the following:

1. Solve the inner optimization problem up to tolerance ε_k. That is, find x_k such that

$$\|\hat{x}(\lambda_k) - x_k\| \le \varepsilon_k$$

2. $$v_Q = \eta \sum_{q=-1}^{Q_k - 1} \prod_{j=Q_k - q}^{Q_k} \left(I - \eta \nabla^2_{xx} \tilde{F}(x_k, \lambda_k; \mathcal{B}_j)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \mathcal{S}_{\nabla_x C}), \tag{3}$$

3. Compute approximate stochastic gradient $\widehat{\nabla} \mathcal{L}(\lambda_k)$ as

$$\widehat{\nabla} \mathcal{L}(\lambda_k) = \nabla_\lambda \tilde{C}(x_k, \lambda_k, \mathcal{S}_{\nabla_\lambda C}) - \nabla^2_{x\lambda} \tilde{F}(x_k, \lambda_k, \mathcal{S}_{\nabla_{x\lambda} F})^\top v_Q \tag{4}$$

4. Update hyperparameters:

$$\lambda_{k+1} = \lambda_k - \nu_k \widehat{\nabla} \mathcal{L}(\lambda_k).$$

3 Stochastic HOAG

3.1 Assumptions

In the following section we adapt the convergence proof of HOAG to the case when all terms are sampled using a single sample.

Assumption 3.1 (Convexity). *The inner function $F(x, \lambda)$ is μ strongly-convex w.r.t. x. For the stochastic setting, the same assumptions hold for $\tilde{F}(x, \lambda; \zeta)$.*

Assumption 3.2 (Smoothness). *Let $(x, \lambda) \in \mathbb{R}^n \times \mathcal{D}$. The loss function $C(x, \lambda)$ and $F(x, \lambda)$ satisfy the following smoothness assumptions:*

- *The function $C(x, \lambda)$ is M-Lipschitz, i.e., for any $(x, \lambda), (x', \lambda') \in \mathbb{R}^n \times \mathcal{D}$,*

$$|C(x, \lambda) - C(x', \lambda')| \leq M \|(x, \lambda) - (x', \lambda')\| .$$

- *$\nabla C(x, \lambda)$ and $\nabla F(x, \lambda)$ are L-Lipschitz, i.e., for any $(x, \lambda), (x', \lambda') \in \mathbb{R}^n \times \mathcal{D}$,*

$$\|\nabla C(x, \lambda) - \nabla C(x', \lambda')\| \leq L \|(x, \lambda) - (x', \lambda')\| ,$$
$$\|\nabla F(x, \lambda) - \nabla F(x', \lambda')\| \leq L \|(x, \lambda) - (x', \lambda')\| .$$

For the stochastic case, the same assumptions hold for $F(x, \lambda; \xi)$ and $G(x, \lambda, \zeta)$ for any given ξ and ζ.

Assumption 3.3 (Partial Lipschitz Smoothness). *Let $z = (x, \lambda) \in \mathbb{R}^n \times \mathcal{D}$. Suppose the derivatives $\nabla_{x\lambda} F(z)$ and $\nabla_x^2 F(z)$ are τ - and ρ - Lipschitz, i.e., - For any z, z', $\|\nabla_{x\lambda} F(z) - \nabla_{x\lambda} F(z')\| \leq \tau \|z - z'\|$. - For any z, z', $\left\|\nabla_x^2 F(z) - \nabla_y^2 F(z')\right\| \leq \rho \|z - z'\|$. For the stochastic case, the same assumptions hold for $\nabla_{x\lambda} F(z; \zeta)$ and $\nabla_x^2 F(z; \zeta)$ for any ζ.*

Assumption 3.4 (Bounded Gradient). *Assume that the partial gradient $\nabla_{x\lambda}^2 F$ is bounded in norm, i.e. $\|\nabla_{x\lambda}^2 F\| \leq K$.*

Assumption 3.5 (Lower bound on objective). *The sequence of iterates $\{\lambda_k\}$ is contained in an open set over which $\mathcal{L}$ is bounded below by a scalar $\mathcal{L}_{inf}$.*

3.2 Preliminaries

We can obtain the exact hypergradient of the outer problem as

$$\nabla \mathcal{L}(\lambda_k) = \nabla_\lambda C(\hat{x}(\lambda_k), \lambda_k) - \nabla_{x\lambda}^2 F(\hat{x}(\lambda_k), \lambda_k)^\top \left[\nabla_{xx} F(\hat{x}(\lambda_k), \lambda_k)\right]^{-1} \nabla_x C(\hat{x}(\lambda_k), \lambda_k).$$

However, finding the inverse $\left[\nabla_{xx} F(\hat{x}(\lambda_k), \lambda_k)\right]^{-1}$ can be costly, and so we consider an approximation for this term. To be more precise, consider the following:

Lemma 3.6 (Neumann Series). *For non-singular $A \in \mathbb{R}^{n \times n}$,*

$$A^{-1} = \sum_{i=0}^{\infty} (I - A)^i, \quad A \succ 0, \|A\| < 1. \tag{5}$$

Define, for i.i.d. samples ξ_0, ζ_j, for $j = 1, \ldots, Q$,

$$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}\left(x_k, \lambda_k; \zeta_j\right)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \xi_0), \tag{6}$$

where we assume $\prod_{j=Q+1}^{Q}(\cdot) = I$. Choose η, such that $\|I - \eta \nabla_{xx}^2 \tilde{F}\left(x_k, \lambda_k; \zeta_j\right)\| \leq (1 - \eta\mu) < 1$. From this, with (5) we get

$$\mathbb{E}[v_Q] = \eta \sum_{i=0}^{Q} \left[I - \eta \nabla_{xx}^2 F\left(x_k, \lambda_k\right)\right]^i \nabla_x C(x_k, \lambda_k) \approx \left[\nabla_x^2 F\left(x_k, \lambda_k\right)\right]^{-1} \nabla_x C\left(x_k, \lambda_k\right).$$

As such, denote $\mathbb{E}[v_\infty] = \left[\nabla_x^2 F\left(x_k, \lambda_k\right)\right]^{-1} \nabla_x C\left(x_k, \lambda_k\right)$. With this in mind, and the fact that it is not generally feasible to solve the inner problem to full precision, we obtain step 3 of Algorithm 2.

3.3 Case I: Variable Q_k, Constant batch sizes

In order to prove the required result, we could set a variable Neumann series length Q_k. Options considered:

$$\textbf{I)} \quad Q_k = r \log k, \quad \textbf{II)} \quad Q_k = r \log \frac{1}{\epsilon_k}$$

for

$$r > \frac{1}{-\log(1-\eta\mu)}.$$

These are selected, so that the T_5 term is summable. In particular, we have

$$\textbf{I)} \quad T_5 < \frac{\sqrt{2}LM}{\mu} k^{-1}, \quad \textbf{II)} \quad T_5 < \frac{\sqrt{2}LM}{\mu} \epsilon_k.$$

Note that, while the choice of η doesn't affect the summability of T_5, it can be problematic for the Neumann series length Q. That is, if η is close to 0, then r may explode. As such, in practice η should be chosen such that $0 < 1 - \eta\mu < 0.7$.

Theorem 3.7 (Global Convergence (SGD step)). *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5 hold. In Algorithm 2, assume that the stepsize ν_k is chosen such that*

$$\sum_{k=1}^{\infty} \nu_k = \infty, \quad \sum_{k=1}^{\infty} \nu_k^2 < \infty.$$

Suppose η is chosen, such that $\eta\mu < 1$. Choose $Q_k = r \log k$, where $r > \frac{1}{-\log(1-\eta\mu)}$. Assume also, that $\lambda_k \in \mathcal{D}$ for all $k > 0$. Let ϵ_k be such that $\|\hat{x}(\lambda_k) - x_k\| \leq \epsilon_k$. If the sequence ϵ_k obeys

$$\sum_{i=1}^{\infty} \epsilon_i < \infty, \quad \epsilon_k > 0 \quad \forall k \geq 0,$$

then we have

$$\min_{k \leq K} \mathbb{E}\|\nabla \mathcal{L}(\lambda_k)\|^2 \xrightarrow{K \to \infty} 0$$

Proof. An equivalent condition to $\mathcal{L}(\lambda)$ having Lipschitz continuous gradient is that for any $\alpha, \beta \in \mathcal{D}$:

$$\mathcal{L}(\beta) \leq \mathcal{L}(\alpha) + \nabla \mathcal{L}(\alpha)^\top (\beta - \alpha) + \frac{L}{2}\|\beta - \alpha\|^2. \tag{7}$$

Substituting for $\alpha = \lambda_k, \beta = \lambda_{k+1} = \lambda_k - \nu_k \widehat{\nabla} \mathcal{L}(\lambda_k)$,

$$\mathcal{L}(\lambda_{k+1}) \leq \mathcal{L}(\lambda_k) + \nabla \mathcal{L}(\lambda_k)^\top \left(-\nu_k \widehat{\nabla} \mathcal{L}(\lambda_k)\right) + \frac{L}{2}\left\|-\nu_k \widehat{\nabla} \mathcal{L}(\lambda_k)\right\|^2,$$

where $\widehat{\nabla} \mathcal{L}(\lambda_k)$ is the approximate hypergradient of $\mathcal{L}$, evaluated at λ_k, as defined in step 3 of Algorithm 2. Taking expectation, conditioning on λ_k,

$$\mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right] \leq \mathcal{L}(\lambda_k) - \nu_k \nabla \mathcal{L}(\lambda_k)^\top \mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right] + \frac{L\nu_k^2}{2}\mathbb{E}_{\lambda_k}\left[\left\|\widehat{\nabla} \mathcal{L}(\lambda_k)\right\|^2\right].$$

$$= \mathcal{L}(\lambda_k) - \nu_k \left(\nabla \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right)^\top \mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right] - \nu_k \left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\|^2 + \frac{L\nu_k^2}{2}\mathbb{E}_{\lambda_k}\left[\left\|\widehat{\nabla} \mathcal{L}(\lambda_k)\right\|^2\right]$$

$$\leq \mathcal{L}(\lambda_k) + \nu_k \left\|\nabla \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\| \left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\| - \nu_k \left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\|^2 + \frac{L\nu_k^2}{2}\mathbb{E}_{\lambda_k}\left[\left\|\widehat{\nabla} \mathcal{L}(\lambda_k)\right\|^2\right]$$

$$= \mathcal{L}(\lambda_k) + \nu_k \left\|\nabla \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\| \left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla} \mathcal{L}(\lambda_k)\right]\right\| + \frac{L\nu_k^2}{2}\mathrm{Var}(\widehat{\nabla} \mathcal{L}(\lambda_k)) - \left(\nu_k - \frac{L\nu_k^2}{2}\right)\|\mathbb{E}_{\lambda_k}\widehat{\nabla} \mathcal{L}(\lambda_k)\|^2, \tag{8}$$

Let $T_{45} = T_4\epsilon_k + T_5$, where T_4, T_5 are as in (22) of Proposition .6. From Proposition .6 and by assumption on ϵ_k, we have that

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\| \leq T_{45}. \tag{9}$$

Note here, that by the summability of ϵ_k, as well as by the choice of Q_k, T_{45} itself is summable. By reverse triangle inequality, we have that

$$\left|\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\| - \|\nabla\mathcal{L}(\lambda_k)\|\right| \leq T_{45}.$$

Hence,

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\| \leq T_{45} + \|\nabla\mathcal{L}(\lambda_k)\| \quad \text{and} \quad \|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\| \geq \|\nabla\mathcal{L}(\lambda_k)\| - T_{45}.$$

Using the fact that for $x > 0$, $x < \max\{1, x^2\}$ we also have

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\| \leq T_{45} + 1 + \|\nabla\mathcal{L}(\lambda_k)\|^2. \tag{10}$$

Furthermore,

$$\begin{aligned}\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 &\leq T_{45}^2 + 2T_{45}\|\nabla\mathcal{L}(\lambda_k)\| + \|\nabla\mathcal{L}(\lambda_k)\|^2 \leq T_{45}^2 + 2T_{45}\left(1 + \|\nabla\mathcal{L}(\lambda_k)\|^2\right) + \|\nabla\mathcal{L}(\lambda_k)\|^2 \\ &= T_{45}^2 + 2T_{45} + (2T_{45} + 1)\|\nabla\mathcal{L}(\lambda_k)\|^2; \end{aligned}\tag{11}$$

Similarly,

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \geq \|\nabla\mathcal{L}(\lambda_k)\|^2(1 - 2T_{45}) + T_{45}^2 - 2T_{45}. \tag{12}$$

Finally, by proposition .4, we have that

$$\operatorname{Var}\left(\widehat{\nabla}\mathcal{L}(\lambda_k)\right) \leq 2\eta^2 M^2\left(\frac{\kappa^2}{B(2-\eta\mu)^2} + \frac{1}{|\mathcal{S}_{\nabla_x C}|(2\eta\mu - \eta^2\mu^2)}\right)\left(\frac{L^2}{|\mathcal{S}_{\nabla_{x\lambda} F}|} + K^2\right) + \frac{\kappa^2 M^2}{|\mathcal{S}_{\nabla_{x\lambda} F}|} + \frac{M^2}{|\mathcal{S}_{\nabla_\lambda C}|} \leq T_2. \tag{13}$$

Substituting (9),(10), (11), (12) and (13) into (8), and rearranging, we get

$$\begin{aligned}\left(\nu_k - \frac{L\nu_k^2}{2}\right)\left(\|\nabla\mathcal{L}(\lambda_k)\|^2(1 - 2T_{45}) + T_{45}^2 - 2T_{45}\right) - \nu_k T_{45}(\|\nabla\mathcal{L}(\lambda_k)\|^2 + 1 + T_{45}) \leq \\ \leq \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right] + \frac{T_2\nu_k^2}{2L}. \end{aligned}\tag{14}$$

Collecting like terms, and further rearranging, we get

$$\|\nabla\mathcal{L}(\lambda_k)\|^2\left(\underbrace{\nu_k - \frac{L\nu_k^2}{2} - 3\nu_k T_{45} + L\nu_k^2 T_{45}}_{\triangleq\gamma_k}\right) \leq \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right] + \underbrace{\frac{T_2\nu_k^2}{2L} + \frac{L\nu_k^2 T_{45}^2}{2} + 3\nu_k T_{45} - T_{45}L\nu_k^2}_{\triangleq T_{2,k}} \tag{15}$$

Note that, by Cauchy-Schwarz, $\sum_{k=1}^{\infty}\nu_k\epsilon_k \leq \sqrt{\sum_{k=1}^{\infty}\nu_k^2}\sqrt{\sum_{k=1}^{\infty}\epsilon_k^2} < \infty$. Similarly, we have that $\nu_k T_{45}$ is summable. Hence, $T_{2,k}$ is summable. Summing (15) for $k = 1$ to ∞, we get

$$\sum_{k=1}^{\infty}\gamma_k\|\nabla\mathcal{L}(\lambda_k)\|^2 \leq \sum_{k=1}^{\infty}\left(\mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right]\right) + \sum_{k=1}^{\infty}T_{2,k}.$$

Taking total expectation and telescoping, we get

$$\sum_{k=1}^{\infty}\gamma_k\mathbb{E}\|\nabla\mathcal{L}(\lambda_k)\|^2 \leq \mathcal{L}(\lambda_1) - \mathcal{L}_{inf} + \sum_{k=1}^{\infty}T_{2,k}, \tag{16}$$

and so the right-hand side of the inequality (16) is finite. From this, we immediately get that

$$\sum_{k=1}^{\infty}\gamma_k\mathbb{E}\|\nabla\mathcal{L}(\lambda_k)\|^2 < \infty. \tag{17}$$

Furthermore, from (17), we get that

$$\min_{k\leq K}\mathbb{E}\|\nabla\mathcal{L}(\lambda_k)\|^2 \leq \frac{\sum_{k=1}^{K}\gamma_k\mathbb{E}\|\nabla\mathcal{L}(\lambda_k)\|^2}{\sum_{k=1}^{K}\gamma_k} \xrightarrow{K\to\infty} 0$$

□

3.4 Case II: Variable Q_k $\mathcal{O}(1/\epsilon_k)$ batches

4 Discussion

References

[1] Bottou, L., Curtis, F., & Nocedal, J.. (2016). Optimization Methods for Large-Scale Machine Learning.

[2] Ji, K. Yang, J. & Liang. Y. Bilevel optimization: Nonasymptotic analysis and faster algorithms. *International Conference on Machine Learning (ICML)*, 2021.

[3] Pedregosa, F. Hyperparameter optimization with approximate gradient. *Proceedings of The 33rd International Conference on Machine Learning, PMLR* 48:737-746, 2016. Available from https://proceedings.mlr.press/v48/pedregosa16.html.

Appendix

Here will will provide the technical lemmas (with proofs) that are required for the proof of Theorem 3.7. We present an immediate consequence of the assumptions in section 3.1.

Proposition .1 (Bounded variance of $\nabla\tilde{C}, \nabla\tilde{F}, \nabla^2_{x\lambda}\tilde{F}, \nabla^2_{xx}\tilde{F}$. Lemma 1 in [2])**.** *Suppose, Assumption 3.2 holds. Then for any x, λ, ζ,*

$$\begin{aligned}
&\mathbb{E}_\zeta\|\nabla\tilde{C}(x,\lambda,\zeta) - \nabla C(x,\lambda)\|^2 \leq M^2 \\
&\mathbb{E}_\zeta\|\nabla^2_{x\lambda}\tilde{F}(x,\lambda,\zeta) - \nabla^2_{x\lambda}F(x,\lambda)\|^2 \leq L^2 \\
&\mathbb{E}_\zeta\|\nabla^2_{xx}\tilde{F}(x,\lambda,\zeta) - \nabla^2_{xx}F(x,\lambda)\|^2 \leq L^2
\end{aligned}$$

Proposition .2 (Bound on $\|\mathbb{E}v_Q\|$)**.** *Suppose Assumptions 3.1, 3.2 hold. Then*

$$\|\mathbb{E}v_Q\| \leq \frac{M}{\mu}(1-(1-\eta\mu)^{Q+1})$$

Proof of proposition .2.

$$\mathbb{E}v_Q = \|\eta\sum_{i=0}^{Q}(I-\eta\mu\nabla^2_{xx}F)^i\nabla_x C\| \leq \eta\|\sum_{i=0}^{Q}(I-\eta\mu\nabla^2_{xx}F)^i\|\|\nabla_x C\| \leq \eta M\sum_{i=0}^{Q}(1-\eta\mu)^i = \eta M\frac{1-(1-\eta\mu)^{Q+1}}{\eta\mu}$$
$$= \frac{M}{\mu}\left(1-(1-\eta\mu)^{Q+1}\right).$$

□

We now introduce our first major result. The variance of the Q-long Neumann expansion can be controlled by the sampling of $\nabla^2_{xx}F$ and $\nabla_x C$ as follows:

Proposition .3 (Bound on $\mathrm{Var}(v_Q)$)**.** *Suppose Assumptions 3.1, 3.2 and 3.4 hold. Denote the condition number as $\kappa = \frac{L}{\mu}$. Choose η, such that $\eta\mu < 1$. Let $|\mathcal{B}_{Q-1+j}| = B(1-\eta\mu)^{j-1}$. Then we have that*

$$\mathrm{Var}(v_Q) = \mathbb{E}\|v_Q - \mathbb{E}(v_Q)\|^2 \leq \frac{2\eta^2M^2\kappa^2}{B(1-\eta\mu)} + \frac{2\eta M^2}{|\mathcal{S}_{\nabla_x C}|(2-\eta\mu)\mu}.$$

Proof of Proposition .3. Denote

$$\begin{aligned}
A_j &= I - \eta\nabla^2_{xx}\tilde{F}(x_k,\lambda_k,\mathcal{B}_j); \\
\tilde{B} &= \nabla_x\tilde{C}(x_k,\lambda_k,\zeta); \\
A &= \mathbb{E}A_j = I - \eta\nabla^2_{xx}F(x_k,\lambda_k); \\
B &= \mathbb{E}B = \nabla_x C(x_k,\lambda_k);
\end{aligned}$$

Then

$$\mathrm{Var}(v_Q) = \mathbb{E}\|v_Q - \mathbb{E}v_Q\|^2$$

$$= \mathbb{E}\left\|\eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} A_j \tilde{B} - \eta \sum_{i=0}^{Q} A^i B\right\|^2$$

$$\leq 2\mathbb{E}\left\|\eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} A_j \tilde{B} - \eta \sum_{i=0}^{Q} A^i \tilde{B}\right\|^2 + 2\mathbb{E}\left\|\eta \sum_{i=0}^{Q} A^i \tilde{B} - \eta \sum_{i=0}^{Q} A^i B\right\|^2$$

$$= 2\mathbb{E}\left\|\eta \sum_{q=-1}^{Q-1} \left(\prod_{j=Q-q}^{Q} A_j - A^{q+1}\right) \tilde{B}\right\|^2 + 2\mathbb{E}\left\|\eta \sum_{i=0}^{Q} A^i (\tilde{B} - B)\right\|^2.$$

Now, note that $\mathbb{E}\left(\prod_{j=Q-q}^{Q} A_j - A^{q+1}\right) = 0$, and that each A_i is independently sampled. Expanding the first term, we get

$$2\mathbb{E}\left\|\eta \sum_{q=-1}^{Q-1} \left(\prod_{j=Q-q}^{Q} A_j - A^{q+1}\right) \tilde{B}\right\|^2 + 2\mathbb{E}\left\|\eta \sum_{i=0}^{Q} A^i (\tilde{B} - B)\right\|^2.$$

$$\leq 2\eta^2 \sum_{q=-1}^{Q-1} \mathbb{E}\left\|\left(\prod_{j=Q-q}^{Q} A_j - A^{q+1}\right) \tilde{B}\right\|^2 + 2\eta^2 \left\|\sum_{i=0}^{Q} A^i\right\|^2 \mathbb{E}\|\tilde{B} - B\|^2$$

$$\leq 2\eta^2 \sum_{q=-1}^{Q-1} \mathbb{E}\left\|\prod_{j=Q-q}^{Q} A_j - A^{q+1}\right\|^2 \mathbb{E}\left\|\tilde{B}\right\|^2 + 2\eta^2 \sum_{i=0}^{Q} \|A\|^{2i} \mathbb{E}\|\tilde{B} - B\|^2$$

$$= 2\eta^2 \sum_{q=0}^{Q} \mathbb{E}\left\|\prod_{j=Q-q+1}^{Q} A_j - A^{q}\right\|^2 \mathbb{E}\left\|\tilde{B}\right\|^2 + 2\eta^2 \frac{1 - \|A\|^{2Q}}{1 - \|A\|^2} \mathbb{E}\|\tilde{B} - B\|^2$$

We will now bound $\mathbb{E}M_i$ for $M_i = \left\|\prod_{j=Q-i+1}^{Q} A_j - A^i\right\|^2$. Note, that $M_0 = 0$. As in the proof of proposition 3 in [2], we write

$$\prod_{j=Q-q+1}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right) = \prod_{j=Q-q+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right) - \eta \nabla_x^2 \tilde{F}_j \prod_{j=Q-q+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right)$$

Then, we have (denoting $\tilde{F}\left(x_k, \lambda_k; \mathcal{B}_j\right) = \tilde{F}_j$)

$$\mathbb{E}M_i = \mathbb{E}\left\|\prod_{j=Q-i+1}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}\left(x_k, \lambda_k; \mathcal{B}_j\right)\right) - \left[I - \eta \nabla_{xx}^2 F\left(x_k, \lambda_k\right)\right]^i\right\|^2$$

$$= \mathbb{E}\left\|\prod_{j=Q-i+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right) - \eta \nabla_x^2 \tilde{F}_{Q-i+1} \prod_{j=Q-i+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right) - \left[I - \eta \nabla_{xx}^2 F\left(x_k, \lambda_k\right)\right]^i\right\|^2$$

Add and subtract $\eta \nabla_x^2 F \prod_{j=Q+2-i}(I - \eta \nabla_x^2 \tilde{F}_j)$:

$$\mathbb{E}\left\|\left(\underbrace{(I - \eta \nabla_x^2 F) \prod_{j=Q-i+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right) - \left[I - \eta \nabla_{xx}^2 F\left(x_k, \lambda_k\right)\right]^i}_{c}\right) + \left(\underbrace{(\eta \nabla_x^2 F - \eta \nabla_x^2 \tilde{F}_{Q-i+1}) \prod_{j=Q-i+2}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}_j\right)}_{d}\right)\right\|^2$$

$$= \mathbb{E}\|c\|^2 + \mathbb{E}\|d\|^2 + \underbrace{2\mathbb{E}\langle c, d\rangle}_{=0 \text{ as } \mathbb{E}(\eta\nabla_x^2 F - \eta\nabla_x^2 \tilde{F}_j)=0}$$

Note, that for sample set $\mathcal{B}_{Q-i+1} = \{\xi_j, j = 1, \ldots, |\mathcal{B}_{Q-i+1}|\}$, with i.i.d. ξ_j, by Proposition .1, we obtain

$$\mathbb{E}\|\nabla_x^2 F(x_k, \lambda_k) - \nabla_x^2 \tilde{F}_{Q-i+1}\|^2 = \mathbb{E}\left\|\frac{1}{|\mathcal{B}_{Q-i+1}|}\sum_{j=1}^{|\mathcal{B}_{Q-i+1}|}\left(\nabla_x^2 F(x_k, \lambda_k) - \nabla_x^2 \tilde{F}(x_k, \lambda_k, \xi_j)\right)\right\|^2$$

$$= \frac{1}{|\mathcal{B}_{Q-i+1}|^2}\sum_{j=1}^{|\mathcal{B}_{Q-i+1}|}\mathbb{E}_{\xi_j}\|\nabla_x^2 F(x_k, \lambda_k) - \nabla_x^2 \tilde{F}(x_k, \lambda_k, \xi_j)\|^2 \leq \frac{L^2}{|\mathcal{B}_{Q-i+1}|} \tag{18}$$

Hence, by strong convexity and Cauchy-Schwarz we obtain that

$$\mathbb{E}\|c\|^2 \leq (1-\eta\mu)^2\mathbb{E}M_{i-1}, \quad \mathbb{E}\|d\|^2 \leq \eta^2(1-\eta\mu)^{2i-2}\frac{L^2}{|\mathcal{B}_{Q-i+1}|}.$$

That is,

$$\mathbb{E}M_i \leq (1-\eta\mu)^2\mathbb{E}M_{i-1} + \eta^2(1-\eta\mu)^{2i-2}\frac{L^2}{|\mathcal{B}_{Q-i+1}|}.$$

Telescoping, we get

$$\mathbb{E}M_i \leq (1-\eta\mu)^{2k}\mathbb{E}M_{i-k} + \eta^2 L^2(1-\eta\mu)^{2i-2}\sum_{j=1}^{k}\frac{1}{|\mathcal{B}_{Q-i+j}|}$$

Setting $i = q, k = q$,

$$\mathbb{E}M_q \leq (1-\eta\mu)^{2q}\mathbb{E}M_0 + \eta^2 L^2(1-\eta\mu)^{2q-2}\sum_{j=1}^{q}\frac{1}{|\mathcal{B}_{Q-1+j}|}$$

Note that $\mathbb{E}(M_0) = 0$. Now, setting $|\mathcal{B}_{Q-1+j}| = B(1-\eta\mu)^{j-1}$, we finally get

$$\mathbb{E}M_q \leq \eta^2 L^2(1-\eta\mu)^{2q-2}\sum_{j=1}^{q}\frac{1}{B(1-\eta\mu)^{j-1}} = \frac{\eta^2 L^2(1-\eta\mu)^{2q-2}}{B}\frac{\left(\frac{1}{1-\eta\mu}\right)^q - 1}{\frac{1}{1-\eta\mu} - 1}$$

$$\leq \frac{\eta L^2(1-\eta\mu)^q}{B(1-\eta\mu)\mu}. \tag{19}$$

In a fashion similar to (18), we obtain $\mathbb{E}\|\tilde{B} - B\|^2 \leq \frac{M^2}{|S_{\nabla_x C}|}$. Hence, using the continuity and convexity assumptions, we get

$$\mathrm{Var}(v_Q) \leq 2\eta^2\sum_{q=0}^{Q}\mathbb{E}M_q\mathbb{E}\left\|\tilde{B}\right\|^2 + 2\eta^2\frac{1-\|A\|^{2Q}}{1-\|A\|^2}\mathbb{E}\|\tilde{B} - B\|^2$$

$$\leq 2\eta^2 M^2\left(\sum_{q=0}^{Q}\frac{\eta L^2(1-\eta\mu)^q}{B(1-\eta\mu)\mu} + \frac{1}{|S_{\nabla_x C}|}\frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}\right).$$

Using the fact, that $\sum_{q=0}^{N} x^q \leq \frac{1}{1-x}$, we get

$$\mathrm{Var}(v_Q) \leq 2\eta^2 M^2\left(\frac{L^2}{B(1-\eta\mu)\mu^2} + \frac{1}{|S_{\nabla_x C}|}\frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}\right)$$

Finally, taking into account, that $\kappa = \frac{L}{\mu}$, we get

$$\mathrm{Var}(v_Q) \leq 2\eta^2 M^2\left(\frac{\kappa^2}{B(1-\eta\mu)} + \frac{1}{|S_{\nabla_x C}|}\frac{1}{2\eta\mu - \eta^2\mu^2}\right)$$

□

Recall, that

$$\widehat{\nabla}\mathcal{L}(\lambda_k) = \nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1) - \nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q.$$

Proposition .4 (Bound on $\mathrm{Var}(\widehat{\nabla}\mathcal{L})$)**.** *Suppose Assumptions 3.1, 3.2 and 3.4 hold. Choose η, such that $\eta\mu < 1$. Denote the condition number as $\kappa = \frac{L}{\mu}$. Let Let $|\mathcal{B}_{Q-1+j}| = B(1-\eta\mu)^{j-1}$. Then the variance of the approximate hypergradient satisfies*

$$\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k)) = \mathbb{E}\|\widehat{\nabla}\mathcal{L}(\lambda_k) - \mathbb{E}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \leq$$

$$2\eta^2 M^2\left(\frac{\kappa^2}{B(2-\eta\mu)^2} + \frac{1}{|\mathcal{S}_{\nabla_x C}|(2\eta\mu - \eta^2\mu^2)}\right)\left(\frac{L^2}{|\mathcal{S}_{\nabla_{x\lambda}F}|} + K^2\right) + \frac{\kappa^2 M^2}{|\mathcal{S}_{\nabla_{x\lambda}F}|} + \frac{M^2}{|\mathcal{S}_{\nabla_\lambda C}|}.$$

Proof of Proposition .4.

$$\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k)) = \mathbb{E}\|\widehat{\nabla}\mathcal{L}(\lambda_k) - \mathbb{E}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 =$$

$$\mathbb{E}\|\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1) - \nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q - \left(\nabla_\lambda C(x_k, \lambda_k) - \nabla^2_{x\lambda}F(x_k, \lambda_k)^\top \mathbb{E}v_Q\right)\|^2$$

$$= \mathbb{E}\|\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1) - \nabla_\lambda C(x_k, \lambda_k)\|^2 + \mathbb{E}\|\nabla^2_{x\lambda}F(x_k, \lambda_k)^\top \mathbb{E}v_Q - \nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q\|^2 -$$

$$-2\mathbb{E}\left[\left(\underbrace{\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1) - \nabla_\lambda C(x_k, \lambda_k)}_{\mathbf{0}\text{ in expectation}}\right)^\top \left(\nabla^2_{x\lambda}F(x_k, \lambda_k)^\top \mathbb{E}v_Q - \nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q\right)\right]$$

$$= \mathbb{E}\|\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1) - \nabla_\lambda C(x_k, \lambda_k)\|^2 + \mathbb{E}\|\nabla^2_{x\lambda}F(x_k, \lambda_k)^\top \mathbb{E}v_Q - \nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q\|^2$$

$$= \mathrm{Var}\left(\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1)\right) + \mathrm{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)^\top v_Q\right) \tag{20}$$

$$= \mathrm{Var}\left(\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1)\right) + \mathrm{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)\right)\mathrm{Var}\left(v_Q\right) + \mathrm{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)\right)\|\mathbb{E}[v_Q]\|^2 +$$

$$+\mathrm{Var}\left(v_Q\right)\left\|\mathbb{E}\left[\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)\right]\right\|^2, \tag{21}$$

where we get from (20) to (21) using the the identity $\mathrm{Var}[XY] = \mathrm{Var}[X]\mathrm{Var}[Y] + \mathrm{Var}[X]\mathbb{E}[Y]^2 + \mathrm{Var}[Y]\mathbb{E}[X]^2$ for independent X, Y. Now, by Proposition .1, and Assumption 3.4, we have

$$\mathrm{Var}\left(\nabla_\lambda \tilde{C}(x_k, \lambda_k, \xi_1)\right) \leq \frac{M^2}{|\mathcal{S}_{\nabla_\lambda C}|}, \quad \mathrm{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)\right) \leq \frac{L^2}{|\mathcal{S}_{x\nabla_\lambda F}|}, \quad \left\|\mathbb{E}\left[\nabla^2_{x\lambda}\tilde{F}(x_k, \lambda_k, \xi_2)\right]\right\|^2 \leq K^2.$$

By Proposition .2, we also obtain

$$\|\mathbb{E}[v_Q]\|^2 \leq \frac{M^2}{\mu^2}(1 - (1-\eta\mu)^{Q+1})^2 \leq \frac{M^2}{\mu^2}.$$

Finally, from Proposition .3, we have

$$\mathrm{Var}(v_Q) \leq 2\eta^2 M^2\left(\frac{\kappa^2}{B(2-\eta\mu)^2} + \frac{1}{|\mathcal{S}_{\nabla_x C}|(2\eta\mu - \eta^2\mu^2)}\right).$$

Thus, we can bound (21) by

$$2\eta^2 M^2\left(\frac{\kappa^2}{B(2-\eta\mu)^2} + \frac{1}{|\mathcal{S}_{\nabla_x C}|(2\eta\mu - \eta^2\mu^2)}\right)\left(\frac{L^2}{|\mathcal{S}_{\nabla_{x\lambda}F}|} + K^2\right) + \frac{\kappa^2 M^2}{|\mathcal{S}_{\nabla_{x\lambda}F}|} + \frac{M^2}{|\mathcal{S}_{\nabla_\lambda C}|},$$

and we are done.

$\square$

Proposition .5 (Bound on $\|\mathbb{E}v_Q - \mathbb{E}v_\infty\|$.). *Suppose Assumptions 3.1, 3.2 hold.*

$$\|\mathbb{E}v_Q - \mathbb{E}v_\infty\| \le \frac{M(1-\eta\mu)^{Q+1}}{\mu}.$$

Proof. Proof of proposition .5.

$$\|\mathbb{E}v_Q - \mathbb{E}v_\infty\| = \left\|\eta\sum_{i=0}^{Q}\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^i\nabla_x C(x_k,\lambda_k) - \eta\sum_{i=0}^{\infty}\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^i\nabla_x C(x_k,\lambda_k)\right\|$$

$$= \left\|\eta\sum_{i=Q+1}^{\infty}\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^i\nabla_x C(x_k,\lambda_k)\right\| \le \left\|\eta\sum_{i=Q+1}^{\infty}\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^i\right\|\|\nabla_x C(x_k,\lambda_k)\|$$

$$\le \left\|\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^{Q+1}\right\|\left\|\underbrace{\eta\sum_{i=0}^{\infty}\left[I-\eta\nabla^2_{xx}F\left(x_k,\lambda_k\right)\right]^i}_{[\nabla^2 F(x_k,\lambda_k)]^{-1}}\right\| M$$

Note, that strong convexity of F gives us $[\nabla^2 F]^{-1} \preceq \frac{1}{\mu}I$, and so, we finally get

$$|\mathbb{E}v_Q - \mathbb{E}v_\infty\| \le \frac{(1-\eta\mu)^{Q+1}M}{\mu}.$$

$\square$

Define

$$\widetilde{\nabla}\mathcal{L}(\lambda_k) = \nabla_\lambda C(x_k,\lambda_k) - \nabla^2_{x\lambda}F(x_k,\lambda_k)^\top\left[\nabla_{xx}F(x_k,\lambda_k)\right]^{-1}\nabla_x C(x_k,\lambda_k);$$

We can then bound the difference between our approximate hypergradient $\nabla\mathcal{L}$ and $\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}$ as follows:

Proposition .6 (Bound on $\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|$. Lemma 7 of [2].). *Let*

$$T_4 = \sqrt{2}\left(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\right), \quad T_5 = \sqrt{2}\frac{LM(1-\eta\mu)^Q}{\mu}. \tag{22}$$

Then we have that

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\| \le T_4\epsilon_k + T_5,$$

where ϵ_k is such that $\|\hat{x}(\lambda_k) - x_k\| \le \epsilon_k$.

The above result is based on Proposition .5 and the fact that

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|^2 \le 2\|\widetilde{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|^2 + 2\|\widetilde{\nabla}\mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2.$$