Stochastic Bilevel Optimization

January 18, 2023

1 Problem

We now have the bilevel problem

$$\min_{\lambda \in \mathcal{D} \subset \mathbb{R}^r} \mathcal{L}(\lambda) \triangleq C(\hat{x}(\lambda), \lambda) \tag{1}$$

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

We will denote the sampled terms as follows:

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

2 Preliminaries

Let us list some useful definitions. We have

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

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

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

Definition 2.1 (Jensen's Inequality)**.** *Theorem 1 (Jensen's Inequality) Let φ be a convex function on $\mathbb{R}$ and let $X \in L_1$ be integrable. Then*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Lemma 2.2 (Neumann Series)**.** *For non-singular $A \in \mathbb{R}^{n \times n}$,*

$$A^{-1} = \sum_{i=0}^{\infty} (I - A)^i, \quad A \succ 0, \|A\| < 1. \tag{3}$$

3 Algorithm

Algorithm 1 Stochastic HOAG

1: At iteration $k = 1, 2, \ldots$, given random samples ξ_i, ζ_j, stepsize ν_k, perform the following:

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

$$\mathbb{E}\left[\left\|\hat{x}\left(\lambda_k\right) - x_k\right\|\right] \leq \varepsilon_k$$

2. $$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}\left(x_k, \lambda_k ; \zeta_j\right)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \xi_0), \tag{4}$$

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

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

4. Update hyperparameters:

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

4 Stochastic HOAG

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

Assumption 4.1 (Convexity). *The lower-level function $F(x, \lambda)$ is μ strongly-convex w.r.t. x and the total objective function $\mathcal{L}(\lambda) = C\left(\lambda, \hat{x}(\lambda)\right)$ is nonconvex w.r.t. λ. For the stochastic setting, the same assumptions hold for $F(x, \lambda; \zeta)$ and $\mathcal{L}(\lambda, \zeta)$, respectively.*

Assumption 4.2 (Smoothness). *Let $z = (x, \lambda) \in \mathbb{R}^n \times \mathcal{D}$. The loss function $C(z)$ and $F(z)$ satisfy - The function $C(z)$ is M-Lipschitz, i.e., for any z, z',*

$$\left|C(z) - C\left(z'\right)\right| \leq M\left\|z - z'\right\|.$$

- $\nabla C(z)$ and $\nabla F(z)$ are L-Lipschitz, i.e., for any z, z',

$$\begin{aligned} \left\|\nabla C(z) - \nabla C\left(z'\right)\right\| &\leq L\left\|z - z'\right\|, \\ \left\|\nabla F(z) - \nabla F\left(z'\right)\right\| &\leq L\left\|z - z'\right\|. \end{aligned}$$

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

Assumption 4.3 (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', $\left\|\nabla_{x\lambda} F(z) - \nabla_{x\lambda} F(z')\right\| \leq \tau\left\|z - z'\right\|$. - For any z, z', $\left\|\nabla_x^2 F(z) - \nabla_y^2 F\left(z'\right)\right\| \leq \rho\left\|z - z'\right\|$. 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 4.4 (Bounded Gradient). *Assume that the partial gradient $\nabla_{x\lambda}^2 F$ is bounded in norm, i.e. $\|\nabla_{x\lambda}^2 F\| \leq K$.*

4.1 Convergence Proof Attempt

Firstly, we will present an immediate consequence of the assumptions in the previous section.

Proposition 4.5 (Bounded variance of $\nabla \tilde{C}, \nabla \tilde{F}, \nabla_{x\lambda}^2 \tilde{F}, \nabla_{xx}^2 \tilde{F}$. Lemma 1 in [2]). *Suppose, Assumption 4.2 holds. Then for any $z = (x, \lambda), \zeta$,*

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

Note that

$$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \left(I - \eta \nabla^2_{xx} \tilde{F}\left(x_k, \lambda_k; \zeta_j\right)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \xi_0), \tag{5}$$

where we assume $\prod_{j=Q+1}^{Q}(\cdot) = I$. From this we easily get

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

Denote by $\mathbb{E}[v_\infty]$:

$$\mathbb{E}[v_\infty] = \left[\nabla^2_x F\left(x_k, \lambda_k\right)\right]^{-1} \nabla_x C\left(x_k, \lambda_k\right)$$

Proposition 4.6 (Bound on $\|\mathbb{E}v_Q\|$)**.** *Suppose Assumptions 4.1, 4.2 hold. Then*

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

Proposition 4.7 (Bound on $\mathrm{Var}(v_Q)$)**.** *Suppose Assumptions 4.1, 4.2 hold. Choose η, such that $\eta\mu < 1$. Then we have that*

$$\mathrm{Var}(v_Q) = \mathbb{E}\|v_Q - \mathbb{E}(v_Q)\|^2 \leq \frac{2\eta^3 M^2 L^2}{\mu}\left(\frac{1-(1-\eta\mu)^{2Q+2}}{1-(1-\eta\mu)^2} - (1-\eta\mu)^2 \frac{1-(1-\eta\mu)^{3Q+3}}{1-(1-\eta\mu)^3}\right) + 2\eta^2 M^2 \frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}.$$

Proposition 4.8 (Bound on $\mathrm{Var}(\widehat{\nabla}\mathcal{L})$)**.** *Suppose Assumptions 4.1, 4.2 hold. Then the variance of the approximate hypergradient satisfies*

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

$$\mathrm{Var}(v_Q)(L^2 + K^2) + L^2\|\mathbb{E}v_Q\|^2 + M^2 \leq$$

$$\left(\frac{2\eta^3 M^2 L^2}{\mu}\left(\frac{1-(1-\eta\mu)^{2Q+2}}{1-(1-\eta\mu)^2} - (1-\eta\mu)^2 \frac{1-(1-\eta\mu)^{3Q+3}}{1-(1-\eta\mu)^3}\right) + 2\eta^2 M^2 \frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}\right)(L^2 + K^2) +$$

$$\frac{L^2 M^2}{\mu^2}\left(1 - (1-\eta\mu)^{Q+1}\right)^2 + M^2.$$

In particular, we have that if $\mathrm{Var}(v_Q)$ is bounded, then so is $\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k))$.

Theorem 4.9 (Global Convergence (SGD step))**.** *In Algorithm 3, assume that the stepsize ν_k is chosen such that*

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

Assume also, that $\lambda_k \in \mathcal{D}$ for all $k > 0$. 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

$$\min_{K \leq k} \mathbb{E}\left[\|\nabla\mathcal{L}(\lambda_K)\|\right] \xrightarrow{k\to\infty} 0.$$

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

$$\mathcal{L}(\beta) \leq \mathcal{L}(\alpha) + \nabla\mathcal{L}(\alpha)^\top(\beta - \alpha) + \frac{L}{2}\|\beta - \alpha\|^2. \tag{6}$$

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

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

Taking expectation, conditioning on λ_k,

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

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

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

$$= \mathcal{L}(\lambda_k) + \frac{\nu_k}{L}\left\|\nabla\mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\|\left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\| + \frac{\nu_k^2}{2L}\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k)) - \underbrace{\left(\frac{\nu_k}{L} - \frac{\nu_k^2}{2L}\right)}_{\geq 0}\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2,$$

Rearranging, we get

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

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

By Assumptions 4.1, 4.2, 4.4, also assuming that $0 < \eta\mu \leq 1$ we can bound $\left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\|$ as

$$\left\|\mathbb{E}_{\lambda_k}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\| \leq M + \frac{K}{\mu}M(1-(1-\eta\mu)^{Q+1}) \leq M + \frac{K}{\mu}M.$$

Furthermore, it can be shown (Lemma 7 of [2]) that

$$\left\|\nabla\mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\| = \mathcal{O}\left(\|x_k - \hat{x}(\lambda_k)\|\right) = \mathcal{O}(\epsilon_k).$$

We now show that $\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k)$ is summable with respect to k.

4.2 Rest of original proof (ignore)

Now, by Theorem **??**, $\|\nabla\mathcal{L}(\lambda_k) - \mathbb{E}\widehat{\nabla}\mathcal{L}(\lambda_k)\| = \mathcal{O}(\epsilon_k)$. As $\mathcal{D}$ is bounded (by Heine-Borel), we have that $\exists M > 0$, such that

$$\|\nabla\mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|\|\lambda_k - \lambda_{k+1}\| < M\epsilon_k.$$

Applying this to (**??**), we get

$$\mathcal{L}(\lambda_{k+1}) \leq \mathcal{L}(\lambda_k) + M\epsilon_k - \frac{1}{2L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2. \tag{7}$$

Rewriting, we get

$$\frac{1}{2L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \leq \mathcal{L}(\lambda_k) - \mathcal{L}(\lambda_{k+1}) + M\epsilon_k. \tag{8}$$

By the extreme-value theorem, since L is defined on a compact set and has continuous derivatives, it has a lower bound K. Thus, taking the sum of (8) for $k = m$ to ∞, we get

$$\frac{1}{2L}\sum_{k=m}^{\infty}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \leq \mathcal{L}(\lambda_m) - K + M\sum_{k=m}^{\infty}\epsilon_k. \tag{9}$$

$\{\epsilon_k\}_{k=1}^{\infty}$ is summable by assumption, thus $\sum_{k=m}^{\infty}\epsilon_k < \infty$, $\epsilon_k \to 0$ and the RHS of (9) is finite. Hence the LHS of (9) must also be finite. Hence $\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \to 0 \iff \|\widehat{\nabla}\mathcal{L}(\lambda_k)\| \to 0$. Recall, that $\|\nabla\mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\| = \mathcal{O}(\epsilon_k)$, hence

$$\|\nabla\mathcal{L}(\lambda_k)\| \leq \|\widehat{\nabla}\mathcal{L}(\lambda_k)\| + \mathcal{O}(\epsilon_k) \xrightarrow{k\to\infty} 0.$$

□

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

Proof of proposition 4.6.

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

$$= \frac{M}{\mu}\left(1-(1-\eta\mu)^{Q+1}\right).$$

□

Proof of Proposition 4.7. Denote

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

Then

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

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

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

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

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

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

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

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

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

We will now bound $\mathbb{E}M_i$ for $M_i = \left\| \prod_{j=Q-q+1}^{Q} A_j - A^q \right\|^2$. 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

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

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

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

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

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

Using convexity assumptions and Proposition 4.5, we get the bound

$$\mathbb{E}M_i \leq (1 - \eta\mu)^2 \mathbb{E}M_{i-1} + \eta^2 (1 - \eta\mu)^{2q-2} L^2.$$

Telescoping, we get

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

Setting $i = q, k = q$,

$$\mathbb{E}M_q \leq (1 - \eta\mu)^{2q-2} \mathbb{E}M_0 + \eta^2 L^2 (1 - \eta\mu)^{2q-2} \sum_{j=1}^{q} (1 - \eta\mu)^{j-1}$$

Note that $\mathbb{E}(M_0) = 0$. Thus, we finally get

$$\mathbb{E}M_q \leq \eta^2 L^2 (1 - \eta\mu)^{2q-2} \sum_{j=0}^{q} (1 - \eta\mu)^j = \eta^2 L^2 (1 - \eta\mu)^{2q-2} \frac{1 - (1 - \eta\mu)^{q+1}}{1 - (1 - \eta\mu)} = \frac{\eta L^2}{\mu} \left((1 - \eta\mu)^{2q-2} - (1 - \eta\mu)^{3q-1} \right) \tag{10}$$

Hence,

$$\mathbb{E}\left\|\prod_{j=Q-q+1}^{Q} A_j - A^q\right\|^2 \leq \eta^2 L^2 (1-\eta\mu)^{2q-2} \frac{1-(1-\eta\mu)^{q+1}}{1-(1-\eta\mu)} = \frac{\eta L^2}{\mu}\left((1-\eta\mu)^{2q-2} - (1-\eta\mu)^{3q-1}\right)$$

Hence, using the continuity and convexity assumptions, we get

$$\begin{aligned}
\mathrm{Var}(v_Q) &\leq 2\eta^2 \sum_{q=0}^{Q} \mathbb{E}\left\|\prod_{j=Q-q+1}^{Q} A_j - A^q\right\|^2 \mathbb{E}\left\|\tilde{B}\right\|^2 + 2\eta^2 \frac{1-\|A\|^{2Q}}{1-\|A\|^2}\mathbb{E}\|\tilde{B}-B\|^2 \\
&\leq 2\eta^2 M^2 \left(\sum_{q=0}^{Q} \frac{\eta L^2}{\mu}\left((1-\eta\mu)^{2q-2} - (1-\eta\mu)^{3q-1}\right) + \frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}\right) \\
&= \frac{2\eta^3 M^2 L^2}{\mu}\left(\frac{1-(1-\eta\mu)^{2Q+2}}{1-(1-\eta\mu)^2} - (1-\eta\mu)^2 \frac{1-(1-\eta\mu)^{3Q+3}}{1-(1-\eta\mu)^3}\right) + 2\eta^2 M^2 \frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}
\end{aligned}$$

□

Proof of Proposition 4.8.

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

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

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

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

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

$$= \mathrm{Var}\left(\nabla_\lambda \tilde{C}(x_k,\lambda_k,\xi_1)\right) + \mathrm{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k,\lambda_k,\xi_2)^\top v_Q\right) \tag{11}$$

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

$$+\mathrm{Var}\left(v_Q\right)\left\|\mathbb{E}\left[\nabla^2_{x\lambda}\tilde{F}(x_k,\lambda_k,\xi_2)\right]\right\|^2, \tag{12}$$

where we get from (11) to (12) 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 4.5, and Assumption 4.4, we have

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

By Proposition 4.6, we also obtain

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

Finally, from Proposition 4.7, we have

$$\mathrm{Var}(v_Q) \leq \frac{2\eta^3 M^2 L^2}{\mu}\left(\frac{1-(1-\eta\mu)^{2Q+2}}{1-(1-\eta\mu)^2} - (1-\eta\mu)^2 \frac{1-(1-\eta\mu)^{3Q+3}}{1-(1-\eta\mu)^3}\right) + 2\eta^2 M^2 \frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}.$$

Thus, we can bound (12) by

$$\left(\frac{2\eta^3M^2L^2}{\mu}\left(\frac{1-(1-\eta\mu)^{2Q+2}}{1-(1-\eta\mu)^2}-(1-\eta\mu)^2\frac{1-(1-\eta\mu)^{3Q+3}}{1-(1-\eta\mu)^3}\right)+2\eta^2M^2\frac{1-(1-\eta\mu)^{2Q}}{1-(1-\eta\mu)^2}\right)\left(L^2+K^2\right)+$$

$$M^2+\frac{L^2M^2}{\mu^2}(1-(1-\eta\mu)^{Q+1})^2,$$

and we are done.

□