Stochastic Bilevel Optimization

January 31, 2023

1 Problem

We now have the bilevel problem

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

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

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 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

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

2.
$$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}(x_k, \lambda_k; \zeta_j)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \xi_0), \tag{3}$$

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

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

4. Update hyperparameters:

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

3 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 3.1 (Convexity). *The lower-level function $F(x, \lambda)$ is μ strongly-convex w.r.t. x and the total objective function $\mathcal{L}(\lambda) = C(\lambda, \hat{x}(\lambda))$ 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 3.2 (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',*

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

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

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

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

Assumption 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 $\nabla_x^2F(z)$ 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', $\left\|\nabla_x^2F(z)-\nabla_y^2F(z')\right\|\leq\rho\|z-z'\|$. For the stochastic case, the same assumptions hold for $\nabla_{x\lambda}F(z;\zeta)$ and $\nabla_x^2F(z;\zeta)$ for any ζ.*

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

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 $\mathcal{L}_{inf}$.*

3.1 Preliminary Results

Let us list some useful definitions. We have

$$\begin{aligned}\nabla\mathcal{L}(\lambda_k)&=\nabla_\lambda C(\hat{x}(\lambda_k),\lambda_k)-\nabla_{x\lambda}^2F(\hat{x}(\lambda_k),\lambda_k)^\top\left[\nabla_{xx}F(\hat{x}(\lambda_k),\lambda_k)\right]^{-1}\nabla_x C(\hat{x}(\lambda_k),\lambda_k)\\ \widetilde{\nabla}\mathcal{L}(\lambda_k)&=\nabla_\lambda C(x_k,\lambda_k)-\nabla_{x\lambda}^2F(x_k,\lambda_k)^\top\left[\nabla_{xx}F(x_k,\lambda_k)\right]^{-1}\nabla_x C(x_k,\lambda_k)\\ \widehat{\nabla}\mathcal{L}(\lambda_k)&=\nabla_\lambda\tilde{C}(x_k,\lambda_k,\xi_1)-\nabla_{x\lambda}^2\tilde{F}(x_k,\lambda_k,\xi_2)^\top v_Q\end{aligned}$$

Definition 3.6 (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 3.7 (Neumann Series)**.** *For non-singular $A\in\mathbb{R}^{n\times n}$,*

$$A^{-1}=\sum_{i=0}^{\infty}(I-A)^i,\quad A\succ 0,\|A\|<1. \tag{4}$$

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

Proposition 3.8 (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 3.2 holds. Then for any $z=(x,\lambda),\zeta$,*

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

Note that

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

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

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

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

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

Proposition 3.9 (Bound on $\|\mathbb{E}v_Q\|$)**.** *Suppose Assumptions 3.1, 3.2 hold. Then*

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

Proposition 3.10 (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 $\eta\mu < 1$. Then we have that*

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

Proposition 3.11 (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 $\kappa = \frac{L}{\mu}$. Then the variance of the approximate hypergradient satisfies*

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

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

$$\left(2\eta^3 M^2 L\kappa + \frac{2\eta M^2}{\mu}\right)\left(L^2+K^2\right) + \kappa^2 M^2 + M^2.$$

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

Proposition 3.12 (Bound on $\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|$. Lemma 7 of [2].)**.** *Let*

$$T_4 = \sqrt{2}\left(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\right), \quad T_5 = \sqrt{2}\frac{LM(1-\eta\mu)^Q}{\mu}. \tag{6}$$

Then we have that

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|_2 \leq T_4\epsilon_k + T_5.$$

3.2 Convergence Proof

Theorem 3.13 (Global Convergence (SGD step))**.** *In Algorithm 2, assume that the stepsize ν_k is chosen such that*

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

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

$$\liminf_{k\to\infty}\mathbb{E}\left[\|\nabla\mathcal{L}(\lambda_k)\|_2^2\right] = 0.$$

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

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

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 (6) of Proposition 3.12. From Proposition 3.12 and by assumption on ϵ_k, we have that

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k) - \nabla\mathcal{L}(\lambda_k)\|_2 \leq T_{45}. \tag{9}$$

By reverse triangle inequality, we have that

$$\left|\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|_2 - \|\nabla\mathcal{L}(\lambda_k)\|_2\right| \leq T_{45}.$$

Hence,

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|_2 \leq T_{45} + \|\nabla\mathcal{L}(\lambda_k)\|_2$$

Note, that from this we have

$$\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|_2 \leq T_{45} + 1 + \|\nabla\mathcal{L}(\lambda_k)\|_2^2 \tag{10}$$

Furthermore,

$$\begin{aligned}\|\mathbb{E}_{\lambda_k}\widehat{\nabla}\mathcal{L}(\lambda_k)\|_2^2 &\leq T_{45}^2 + 2T_{45}\|\nabla\mathcal{L}(\lambda_k)\|_2 + \|\nabla\mathcal{L}(\lambda_k)\|_2^2 \leq T_{45}^2 + 2T_{45}\left(1 + \|\nabla\mathcal{L}(\lambda_k)\|_2^2\right) + \|\nabla\mathcal{L}(\lambda_k)\|_2^2 \\ &= T_{45}^2 + 2T_{45} + (2T_{45}+1)\|\nabla\mathcal{L}(\lambda_k)\|_2^2. \end{aligned} \tag{11}$$

Finally, by proposition 3.11, we have that

$$\mathrm{Var}\left(\widehat{\nabla}\mathcal{L}(\lambda_k)\right) \leq T_2 = \left(2\eta^3 M^2 L\kappa + \frac{2\eta M^2}{\mu}\right)\left(L^2 + K^2\right) + \kappa^2 M^2 + M^2. \tag{12}$$

Set $M_1 = T_{45}^2 + 2T_{45}$, $M_2 = 2T_{45} + 1$. Substituting (9),(10), (11) and (12) into (8), and rearranging, we get

$$\begin{aligned}\left(\nu_k - \frac{L\nu_k^2}{2}\right)(M_1 + M_2\|\nabla\mathcal{L}(\lambda_k)\|_2^2) - \nu_k((T_4\epsilon + T_5)\|\nabla\mathcal{L}(\lambda_k)\|_2^2 + T_5^2 + T_5) \leq \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right] + \\ + \nu_k\epsilon_k\left(\underbrace{T_4^2 T_\epsilon + 2T_4T_5 + T_4}_{\triangleq T_6}\right) + \frac{T_2\nu_k^2}{2L}\end{aligned} \tag{13}$$

Now, let us inspect the left-hand-side of (13). Expanding M_1, M_2 and collecting like terms, we get

$$\left(\nu_k - \frac{L\nu_k^2}{2}\right)(M_1 + M_2\|\nabla\mathcal{L}(\lambda_k)\|_2^2) - \nu_k((T_4\epsilon + T_5)\|\nabla\mathcal{L}(\lambda_k)\|_2^2 + T_5^2 + T_5) =$$

$$= \left(\underbrace{\nu_k(\underbrace{2T_{45} + 1 - T_5}_{\geq 0}) - \frac{M_2 L\nu_k^2}{2} - T_4\nu_k\epsilon_k}_{\triangleq \gamma_k}\right)\|\nabla\mathcal{L}(\lambda_k)\|_2^2 + \nu_k\epsilon_k(T_4\epsilon + 2T_4T_5 + 2T_4) + \nu_k T_5 \geq$$

$$\geq \gamma_k\|\nabla\mathcal{L}(\lambda_k)\|_2^2 + \nu_k\epsilon_k(T_4\epsilon + 2T_4T_5 + 2T_4).$$

We can now rewrite the inequality (13) as

$$\gamma_k \|\nabla \mathcal{L}(\lambda_k)\|_2^2 \leq \mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right] + \frac{T_2 \nu_k^2}{2L} + \nu_k \epsilon_k \underbrace{(T_6 - T_4\epsilon - 2T_4T_5 - 2T_4)}_{\triangleq T_7} \tag{14}$$

Note that, by Cauchy-Schwarz, $\sum_{k=1}^{\infty} \nu_k \epsilon_k \leq \sqrt{\sum_{k=1}^{\infty} \nu_k^2}\sqrt{\sum_{k=1}^{\infty} \epsilon_k^2} < \infty$. As such, γ_k is summable. Summing (14) for $k = 1$ to ∞, we get

$$\sum_{k=1}^{\infty} \gamma_k \|\nabla \mathcal{L}(\lambda_k)\|_2^2 \leq \sum_{k=1}^{\infty} \left(\mathcal{L}(\lambda_k) - \mathbb{E}_{\lambda_k}\left[\mathcal{L}(\lambda_{k+1})\right]\right) + \frac{T_2}{2L}\sum_{k=1}^{\infty} \nu_k^2 + T_7 \sum_{k=1}^{\infty} \nu_k \epsilon_k.$$

Taking total expectation and telescoping, we get

$$\sum_{k=1}^{\infty} \gamma_k \mathbb{E}\|\nabla \mathcal{L}(\lambda_k)\|_2^2 \leq \mathcal{L}(\lambda_1) - \mathcal{L}_{inf} + \frac{T_2}{2L}\sum_{k=1}^{\infty} \nu_k^2 + T_7 \sum_{k=1}^{\infty} \nu_k \epsilon_k. \tag{15}$$

and so the right-hand side of the inequality (15) is finite. From this, we immediately get that

$$\sum_{k=1}^{\infty} \gamma_k \mathbb{E}\|\nabla \mathcal{L}(\lambda_k)\|_2^2 < \infty. \tag{16}$$

Furthermore, from (16), we get that

$$\liminf_{k\to\infty} \mathbb{E}\|\nabla \mathcal{L}(\lambda_k)\|_2^2 = 0.$$

□

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

Proof of proposition 3.9.

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

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

□

Proof of Proposition 3.10. Denote

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

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 $M_i = \left\|\prod_{j=Q-q+1}^{Q} A_j - A^q\right\|^2$. 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 3.8, 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

$$\mathbb{E}M_q \leq \eta^2 L^2(1-\eta\mu)^{2q-2}\sum_{j=0}^{q}(1-\eta\mu)^{j} = \eta^2 L^2(1-\eta\mu)^{2q-2}\frac{1-(1-\eta\mu)^{q+1}}{1-(1-\eta\mu)} = \frac{\eta L^2}{\mu}\left((1-\eta\mu)^{2q-2} - (1-\eta\mu)^{3q-1}\right) \tag{17}$$

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

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

Furthermore, since $\eta\mu < 1$, we have that

$$\frac{2\eta^3 M^2 L^2 + 2\eta M^2}{2\mu - \eta\mu^2} \leq \frac{2\eta^3 M^2 L^2 + 2\eta M^2}{2\mu - \mu} = 2\eta^3 M^2 L\kappa + \frac{2\eta M^2}{\mu}$$

$\square$

Proof of Proposition 3.11.

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

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

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

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

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

$$= \operatorname{Var}\left(\nabla_\lambda \tilde{C}(x_k,\lambda_k,\xi_1)\right) + \operatorname{Var}\left(\nabla^2_{x\lambda}\tilde{F}(x_k,\lambda_k,\xi_2)^\top v_Q\right) \tag{18}$$

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

$$+\operatorname{Var}(v_Q)\left\|\mathbb{E}\left[\nabla^2_{x\lambda}\tilde{F}(x_k,\lambda_k,\xi_2)\right]\right\|^2, \tag{19}$$

where we get from (18) to (19) using the the identity $\operatorname{Var}[XY] = \operatorname{Var}[X]\operatorname{Var}[Y] + \operatorname{Var}[X]\mathbb{E}[Y]^2 + \operatorname{Var}[Y]\mathbb{E}[X]^2$ for independent X, Y. Now, by Proposition 3.8, and Assumption 3.4, we have

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

By Proposition 3.9, we also obtain

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

Finally, from Proposition 3.10, we have

$$\operatorname{Var}(v_Q) \le 2\eta^3 M^2 L\kappa + \frac{2\eta M^2}{\mu}.$$

Thus, we can bound (19) by

$$\left(2\eta^3 M^2 L\kappa + \frac{2\eta M^2}{\mu}\right)\left(L^2 + K^2\right) + \kappa^2 M^2 + M^2,$$

and we are done.

□