Stochastic Bilevel Optimization

January 4, 2023

1 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 1.1 (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*

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

2 Problem

We now have the bilevel problem

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

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

We will denote the sampled terms as follows:

$$\mathbb{E}_\zeta[\nabla^2_{xx} \tilde{F}\left(x_k, \lambda_k; \zeta\right)] = \nabla^2_{xx} F\left(x_k, \lambda_k\right)$$

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

$$\mathbb{E}_\theta[\nabla_\lambda \tilde{C}(x_k, \lambda_k, \theta)] = \nabla_\lambda \tilde{C}(x_k, \lambda_k)$$

$$\mathbb{E}_\kappa[\nabla^2_{x\lambda} \tilde{F}(x_k, \lambda_k, \kappa)] = \nabla^2_{x\lambda} \tilde{F}(x_k, \lambda_k)$$

3 Algorithm

Algorithm 1 Stochastic HOAG

1: At iteration $k = 1, 2, \ldots$, given random samples ξ_i, ζ_j, 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. $$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \left(I - \eta \nabla_{xx}^2 \tilde{F}\left(x_k, \lambda_k; \zeta_j\right)\right) \nabla_x \tilde{C}(x_k, \lambda_k, \xi_0), \tag{3}$$

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:

$$\lambda_{k+1} = P_{\mathcal{D}}\left(\lambda_k - \frac{1}{L}\widehat{\nabla}\mathcal{L}(\lambda_k)\right).$$

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',

$$\left\|\nabla C(z) - \nabla C\left(z'\right)\right\| \leq L\left\|z - z'\right\|,$$
$$\left\|\nabla F(z) - \nabla F\left(z'\right)\right\| \leq L\left\|z - z'\right\|.$$

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.*

4.1 Convergence Proof Attempt

Theorem 4.4 (Bounded Gradient Error (Lemma 7 of [2]))**.** *Suppose Assumptions 4.1, 4.2 and 4.3 hold. Then, conditioning on x_k^D and λ_k, we have*

$$\left\|\mathbb{E}\widehat{\nabla}\mathcal{L}\left(\lambda_k\right) - \nabla\mathcal{L}\left(\lambda_k\right)\right\|^2 \leq 2\left(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\right)^2 \left\|x_k^D - \hat{x}\left(\lambda_k\right)\right\|^2 + \frac{2L^2M^2(1-\eta\mu)^{2Q}}{\mu^2},$$

that is,

$$\left\|\mathbb{E}\widehat{\nabla}\mathcal{L}\left(\lambda_k\right) - \nabla\mathcal{L}\left(\lambda_k\right)\right\| = \mathcal{O}(\epsilon_k), \quad \left\|x_k^D - \hat{x}\left(\lambda_k\right)\right\| \leq \epsilon_k.$$

Theorem 4.5 (Bounded Variance of $\widehat{\nabla}\mathcal{L}$ (Lemma 8 of [2]))**.** *Suppose Assumptions 4.1, 4.2 and 4.3 hold. Assume all sample sizes are 1. Then, we have*

$$\mathbb{E}\left\|\widehat{\nabla}\mathcal{L}\left(\lambda_k\right)-\nabla\mathcal{L}\left(\lambda_k\right)\right\|^2 \leq \frac{4L^2M^2}{\mu^2}+\left(\frac{8L^2}{\mu^2}+2\right)\frac{M^2}{1}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}$$
$$+\left(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\right)^2\mathbb{E}\left\|x_k^D-\hat{x}\left(\lambda_k\right)\right\|^2.$$

That is,

$$\mathbb{E}\left[\left\|\widehat{\nabla}\mathcal{L}\left(\lambda_k\right)-\nabla\mathcal{L}\left(\lambda_k\right)\right\|^2\right]=\mathcal{O}(\mathbb{E}\left[\left\|x_k^D-\hat{x}\left(\lambda_k\right)\right\|^2\right])\stackrel{?}{=}\mathcal{O}(\epsilon^2).$$

Theorem 4.6 (Global Convergence (Simplified) (SGD step))**.** *In Algorithm 3, assume that the update*

$$\lambda_{k+1}=P_{\mathcal{D}}\left(\lambda_k-\frac{1}{L}\widehat{\nabla}\mathcal{L}(\lambda_k)\right),$$

is replaced by SGD with approximate gradient:

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

Assume also, that $\lambda_k\in\mathcal{D}$ for all $k>0$. If the sequence ϵ_k obeys

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

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

$$\mathcal{L}(\beta)\leq\mathcal{L}(\alpha)+\nabla\mathcal{L}(\alpha)^\top(\beta-\alpha)+\frac{L}{2}\|\beta-\alpha\|^2. \tag{4}$$

Substituting for $\alpha=\lambda_k,\beta=\lambda_{k+1}=\lambda_k-\frac{1}{L}\widehat{\nabla}\mathcal{L}(\lambda_k)$, we add and subtract $\widehat{\nabla}\mathcal{L}(\lambda_k)^\top(\lambda_k-\lambda_{k-1})$ to get

$$\mathcal{L}(\lambda_{k+1})\leq\mathcal{L}(\lambda_k)-(\nabla\mathcal{L}(\lambda_k)-\widehat{\nabla}\mathcal{L}(\lambda_k))^\top(\lambda_k-\lambda_{k+1})-\widehat{\nabla}\mathcal{L}(\lambda_k)^\top(\lambda_k-\lambda_{k+1})+\frac{L}{2}\|\lambda_k-\lambda_{k+1}\|^2 \tag{5}$$

$$=\mathcal{L}(\lambda_k)-(\nabla\mathcal{L}(\lambda_k)-\widehat{\nabla}\mathcal{L}(\lambda_k))\top(\lambda_k-\lambda_{k+1})-\frac{1}{L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2+\frac{1}{2L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\quad(\text{using }\lambda_k-\lambda_{k+1}=\frac{1}{L}\widehat{\nabla}\mathcal{L}(\lambda_k)) \tag{6}$$

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

Finally, by Cauchy-Schwarz, we have

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

Taking expectations of both sides, and conditioning on λ_k, we get

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

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

$$\leq\mathcal{L}(\lambda_k)+\frac{1}{L}\sqrt{\mathbb{E}\left[\|\nabla\mathcal{L}(\lambda_k)-\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]}\sqrt{\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]}-\frac{1}{2L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right], \tag{9}$$

where the last line is given by the Cauchy-Schwarz inequality.

4.2 Approach I

That root term looks nasty to work with - let's get rid of it. Denote

$$X = \mathbb{E}\left[\|\nabla \mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right], \quad Y = \mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right].$$

Then we have, that

$$\frac{1}{L}\sqrt{\mathbb{E}\left[\|\nabla \mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]}\sqrt{\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]} - \frac{1}{2L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right] = \frac{1}{L}\sqrt{XY} - \frac{1}{2L}Y$$
$$= \frac{1}{L}X - \frac{1}{4L}Y - \frac{1}{L}\left(\sqrt{X} - \frac{1}{2}\sqrt{Y}\right)^2.$$

Thus, going back to equation (9), we get

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

$$= \mathcal{L}(\lambda_k) + \frac{1}{L}\mathbb{E}\left[\|\nabla \mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right] - \frac{1}{4L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right] - \frac{1}{L}\left(\sqrt{\mathbb{E}\left[\|\nabla \mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]} - \frac{1}{2}\sqrt{\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right]}\right)^2$$
$$\le \mathcal{L}(\lambda_k) + \frac{1}{L}\mathbb{E}\left[\|\nabla \mathcal{L}(\lambda_k) - \widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right] - \frac{1}{4L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right].$$
$$\le \mathcal{L}(\lambda_k) + \frac{M}{L}\mathbb{E}\left[\left\|x_k^D - \hat{x}\left(\lambda_k\right)\right\|^2\right] - \frac{1}{4L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right].$$

where the last line follows from theorem 4.5.Rewriting, we get

$$\frac{1}{4L}\mathbb{E}\left[\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2\right] \le \mathcal{L}(\lambda_k) - \mathbb{E}\mathcal{L}(\lambda_{k+1}) + \frac{M}{L}\mathbb{E}\left[\left\|x_k^D - \hat{x}\left(\lambda_k\right)\right\|^2\right] \tag{10}$$

4.3 Approach II

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

$$\mathcal{L}(\beta) \le \mathcal{L}(\alpha) + \nabla \mathcal{L}(\alpha)^\top(\beta - \alpha) + \frac{L}{2}\|\beta - \alpha\|^2. \tag{11}$$

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

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

Taking expectation, conditioning on λ_k,

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

$$= \mathcal{L}(\lambda_k) - \frac{1}{L}\left(\nabla \mathcal{L}(\lambda_k) - \mathbb{E}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right)^\top\mathbb{E}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right] - \frac{1}{L}\left\|\mathbb{E}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\|^2 + \frac{1}{2L}\mathbb{E}\left[\left\|\widehat{\nabla}\mathcal{L}(\lambda_k)\right\|^2\right]$$

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

$$\le \mathcal{L}(\lambda_k) + \frac{1}{L}\left\|\nabla \mathcal{L}(\lambda_k) - \mathbb{E}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\|\left\|\mathbb{E}\left[\widehat{\nabla}\mathcal{L}(\lambda_k)\right]\right\| + \frac{1}{2L}\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k)) - \frac{1}{2L}\|\mathbb{E}\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2$$

Now, we are interested in controlling $\mathrm{Var}(\widehat{\nabla}\mathcal{L}(\lambda_k))$ through $\mathrm{Var}(v_Q)$, where

$$v_Q = \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} \Big(I - \eta \nabla^2_{xx} \tilde{F}\left(x_k, \lambda_k; \mathcal{B}_j\right)\Big) \nabla_x \tilde{C}(x_k, \lambda_k, \mathcal{S}_C), \tag{12}$$

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

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

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

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

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

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

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

4.4 Rest of original proof

Now, by Theorem 4.4, $\|\nabla\mathcal{L}(\lambda_k) - \mathbb{E}\widehat{\nabla}\mathcal{L}(\lambda_k)\| = \mathcal{O}(\epsilon_k)$. 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 (8), we get

$$\mathcal{L}(\lambda_{k+1}) \leq \mathcal{L}(\lambda_k) + M\epsilon_k - \frac{1}{2L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2. \tag{13}$$

Rewriting, we get

$$\frac{1}{2L}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \leq \mathcal{L}(\lambda_k) - \mathcal{L}(\lambda_{k+1}) + M\epsilon_k. \tag{14}$$

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 (14) for $k = m$ to ∞, we get

$$\frac{1}{2L}\sum_{k=m}^{\infty}\|\widehat{\nabla}\mathcal{L}(\lambda_k)\|^2 \leq \mathcal{L}(\lambda_m) - K + M\sum_{k=m}^{\infty}\epsilon_k. \tag{15}$$

$\{\epsilon_k\}_{k=1}^{\infty}$ is summable by assumption, thus $\sum_{k=m}^{\infty}\epsilon_k < \infty$, $\epsilon_k \to 0$ and the RHS of (15) is finite. Hence the LHS of (15) 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.