CSC/DSCC 574 Continuous Algorithms for Optimization and Sampling — Lecture 6

Stochastic Approximation

Lecturer: Jiaming Liang *February 15, 2024*

1 Stochastic optimization

Sampling from a probability distribution $P_0(x) \propto \exp(-f(x))$ can be cast as an optimization

$$\min_{P \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(P||P_0).$$

Note that $\mathrm{KL}(\cdot||\cdot)$ is not symmetric, $\mathrm{KL}(\mu||\nu) \ge 0$, and $\mathrm{KL}(\mu||\nu) = 0$ if and only if $\mu = \nu$. If we parametrize the target distribution P_0 as P_{θ_0} where $\theta_0 \in \mathbb{R}^n$ and switch P and P_0, then we reformulate the problem as

$$\min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(P_{\theta_0}||P_\theta).$$

By the definition of KL divergence, we have

$$\begin{aligned}
\min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(P_{\theta_0}||P_\theta) &= \min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^d)} \int \log \frac{P_{\theta_0}(z)}{P_\theta(z)} P_{\theta_0}(z) dz \\
&= \min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^d)} \int \log P_{\theta_0}(z) P_{\theta_0}(z) dz - \int \log P_\theta(z) P_{\theta_0}(z) dz \\
&= \int \log P_{\theta_0}(z) P_{\theta_0}(z) dz - \max_{P_\theta \in \mathcal{P}_2(\mathbb{R}^d)} \int \log P_\theta(z) P_{\theta_0}(z) dz \\
&= \int \log P_{\theta_0}(z) P_{\theta_0}(z) dz - \max_{\theta \in \Theta} \mathbb{E}_{z \sim P_{\theta_0}}[\log P_\theta(z)].
\end{aligned}$$

The infinite dimensional optimization problem thus reduces to an n-dimensional problem

$$\max_{\theta \in \Theta} \mathbb{E}_{z \sim P_{\theta_0}}[\log P_\theta(z)], \tag{1}$$

and can be generalized as stochastic optimization (SO)

$$\min_{x \in \mathbb{R}^n} \left\{ \phi(x) := f(x) + h(x) \right\}, \quad f(x) = \mathbb{E}_\xi[F(x,\xi)]. \tag{2}$$

Problem (1) is indeed the maximum likelihood estimation (MLE). A standard way to solve MLE (1) (and SO (2) in general) is to solve its sample average approximation (SAA), namely, taking independent and identically distributed (i.i.d.) samples $Z_1, \ldots, Z_N$ of $Z \sim P_{\theta_0}$ and optimizing the average of the function value samples,

$$\max_{\theta \in \Theta} \left\{ \ell(\theta|Z) := \frac{1}{N} \sum_{i=1}^N \log P_\theta(Z_i) \right\}.$$

Since we take the i.i.d. samples first and then solve the deterministic optimization problem, this is an offline approach. In contrast, we can take a sample of the function value (if necessary) and its first-order information, and perform a (proximal) gradient step. This method is called stochastic approximation (SA) and is an online approach.

2 Stochastic approximation

To study the SA approach for solving (2), we need the following assumptions.

(A1) both f and h are closed and convex functions;

(A2) for almost every $\xi \in \Xi$, a functional oracle $F(\cdot, \xi) : \operatorname{dom} h \to \mathbb{R}$ and a stochastic gradient oracle $s(\cdot, \xi) : \operatorname{dom} h \to \mathbb{R}^n$ satisfying

$$f(x) = \mathbb{E}[F(x, \xi)], \quad f'(x) = \mathbb{E}[s(x, \xi)] \in \partial f(x)$$

for every $x \in \operatorname{dom} h$ are available;

(A3) for every $x \in \operatorname{dom} h$, we have $\mathbb{E}[\|s(x, \xi) - f'(x)\|^2] \leq \sigma^2$;

(A4) for every $x, y \in \operatorname{dom} h$,

$$f(x) - f(y) - \langle f'(y), x - y \rangle \leq 2M\|x - y\| + \frac{L}{2}\|x - y\|^2. \tag{3}$$

In this section, we are particularly interested in the stochastic version of the proximal subgradient method, which is an SA-type method.

Algorithm 1 Stochastic subgradient method

Input: Initial point $x_0 \in \mathbb{R}^n$

for $k \geq 0$ **do**

Step 1. Choose $\lambda_k \in (0, 1/(2L))$ and generate a stochastic gradient $s(x_k; \xi_k)$

Step 2. Compute

$$x_{k+1} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \langle s(x_k; \xi_k), u \rangle + h(u) + \frac{1}{2\lambda_k}\|u - x_k\|^2 \right\}.$$

end for

2.1 Convergence in expectation

Lemma 1. *For every $k \geq 0$, we have*

$$\begin{aligned} \lambda_k \left(\phi(x_{k+1}) - \phi(x_*) \right) \leq \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\|x_{k+1} - x_*\|^2 + \frac{4\lambda_k^2 M^2}{1 - 2\lambda_k L} \\ + \lambda_k \langle s(x_k; \xi_k) - f'(x_k), x_* - x_k \rangle + \lambda_k^2 \|s(x_k; \xi_k) - f'(x_k)\|^2. \end{aligned} \tag{4}$$

Proof. It follows from step 2 of Algorithm 1 that for every $u \in \operatorname{dom} h$,

$$\langle s(x_k;\xi_k),u\rangle+h(u)+\frac{1}{2\lambda_k}\|u-x_k\|^2 \geq \langle s(x_k;\xi_k),x_{k+1}\rangle+h(x_{k+1})+\frac{1}{2\lambda_k}\|x_{k+1}-x_k\|^2+\frac{1}{2\lambda_k}\|x_{k+1}-u\|^2.$$

Taking $u = x_*$ in the above inequality and using the convexity of f and (3) with $(x,y) = (x_{k+1}, x_k)$, we have

$$\begin{aligned}
&f(x_*) - \langle f'(x_k), x_* - x_k\rangle + h(x_*) + \langle s(x_k;\xi_k), x_*\rangle + \frac{1}{2\lambda_k}\|x_k - x_*\|^2 \\
\geq& f(x_k) + h(x_*) + \langle s(x_k;\xi_k), x_*\rangle + \frac{1}{2\lambda_k}\|x_k - x_*\|^2 \\
\geq& f(x_k) + h(x_{k+1}) + \langle s(x_k;\xi_k), x_{k+1}\rangle + \frac{1}{2\lambda_k}\|x_{k+1} - x_k\|^2 + \frac{1}{2\lambda_k}\|x_{k+1} - x_*\|^2 \\
\geq& f(x_{k+1}) - \langle f'(x_k), x_{k+1} - x_k\rangle - 2M\|x_{k+1} - x_k\| + \frac{1-\lambda_k L}{2\lambda_k}\|x_{k+1} - x_k\|^2 \\
&+ h(x_{k+1}) + \langle s(x_k;\xi_k), x_{k+1}\rangle + \frac{1}{2\lambda_k}\|x_{k+1} - x_*\|^2.
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
\lambda_k\left(\phi(x_{k+1}) - \phi(x_*)\right) \leq& \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\|x_{k+1} - x_*\|^2 + 2\lambda_k M\|x_{k+1} - x_k\| - \frac{1-\lambda_k L}{2}\|x_{k+1} - x_k\|^2 \\
&+ \lambda_k\langle s(x_k;\xi_k) - f'(x_k), x_* - x_k\rangle + +\lambda_k\langle s(x_k;\xi_k) - f'(x_k), x_k - x_{k+1}\rangle.
\end{aligned}$$

Using the above inequality, the Cauchy-Schwarz inequality, and the fact $\lambda_k < 1/(2L)$, we have

$$\begin{aligned}
\lambda_k\left(\phi(x_{k+1}) - \phi(x_*)\right) \leq& \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\|x_{k+1} - x_*\|^2 + 2\lambda_k M\|x_{k+1} - x_k\| - \frac{1-\lambda_k L}{2}\|x_{k+1} - x_k\|^2 \\
&+ \lambda_k\langle s(x_k;\xi_k) - f'(x_k), x_* - x_k\rangle + \lambda_k^2\|s(x_k;\xi_k) - f'(x_k)\|^2 + \frac{1}{4}\|x_{k+1} - x_k\|^2 \\
\leq& \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\|x_{k+1} - x_*\|^2 + 2\lambda_k M\|x_{k+1} - x_k\| - \frac{1-2\lambda_k L}{4}\|x_{k+1} - x_k\|^2 \\
&+ \lambda_k\langle s(x_k;\xi_k) - f'(x_k), x_* - x_k\rangle + \lambda_k^2\|s(x_k;\xi_k) - f'(x_k)\|^2.
\end{aligned}$$

Finally, (4) follows from the above inequality and the AM-GM inequality. $\square$

Theorem 1. *If $\lambda_k < 1/(2L)$ for every $k \geq 0$, then*

$$\mathbb{E}_{\xi_{[k-1]}}\left[\phi(\bar{x}_k)\right] - \phi(x_*) \leq \frac{d_0^2 + \sum_{i=0}^{k-1}\frac{8\lambda_i^2 M^2}{1-\lambda_i L} + \sum_{i=0}^{k-1} 2\lambda_i^2\sigma^2}{2\sum_{i=0}^{k-1}\lambda_i} \tag{5}$$

where

$$\bar{x}_k := \frac{\sum_{i=0}^{k-1}\lambda_i x_{i+1}}{\sum_{i=0}^{k-1}\lambda_i}, \quad d_0 := \|x_0 - x_*\|. \tag{6}$$

As a consequence, if

$$\lambda_k = \lambda = \min\left\{\frac{\varepsilon}{16M^2 + 2\sigma^2}, \frac{1}{4L}\right\}, \tag{7}$$

then we find $\bar{x}_k$ such that $\mathbb{E}_{\xi_{[k-1]}}[\phi(\bar{x}_k)] - \phi(x_*) \le \varepsilon$ in at most

$$\max\left\{\frac{4Ld_0^2}{\varepsilon}, \frac{(16M^2 + 2\sigma^2)d_0^2}{\varepsilon^2}\right\}$$

iterations.

Proof. Taking expectation of (4) w.r.t. ξ_k conditioned on $\xi_{[k-1]}$ and using (A2) and (A3), we have

$$\begin{aligned}\lambda_k \mathbb{E}_{\xi_k}\left[\phi(x_{k+1})|\xi_{[k-1]}\right] - \lambda_k \phi(x_*) &\le \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\mathbb{E}_{\xi_k}\left[\|x_{k+1} - x_*\|^2|\xi_{[k-1]}\right] + \frac{4\lambda_k^2 M^2}{1 - 2\lambda_k L} \\ &\quad + \lambda_k^2 \mathbb{E}_{\xi_k}\left[\|s(x_k;\xi_k) - s_k\|^2|\xi_{[k-1]}\right] \\ &\le \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\mathbb{E}_{\xi_k}\left[\|x_{k+1} - x_*\|^2|\xi_{[k-1]}\right] + \frac{4\lambda_k^2 M^2}{1 - 2\lambda_k L} + \lambda_k^2\sigma^2.\end{aligned}$$

Taking expectation of the above inequality w.r.t. $\xi_{[k-1]}$ and using the law of total expectation, we have

$$\lambda_k \mathbb{E}_{\xi_{[k]}}[\phi(x_{k+1})] - \lambda_k\phi(x_*) \le \frac{1}{2}\mathbb{E}_{\xi_{[k-1]}}\left[\|x_k - x_*\|^2\right] - \frac{1}{2}\mathbb{E}_{\xi_{[k]}}\left[\|x_{k+1} - x_*\|^2\right] + \frac{4\lambda_k^2 M^2}{1 - 2\lambda_k L} + \lambda_k^2\sigma^2.$$

Summing the above inequality from $k = 0$ to $k - 1$, we obtain

$$\sum_{i=0}^{k-1} \lambda_i \left[\mathbb{E}_{\xi_{[i]}}[\phi(x_{i+1})] - \phi(x_*)\right] \le \frac{1}{2}d_0^2 + \sum_{i=0}^{k-1}\frac{4\lambda_i^2 M^2}{1 - 2\lambda_i L} + \sum_{i=0}^{k-1}\lambda_i^2\sigma^2.$$

Using the convexity of ϕ and the definition of $\bar{x}_k$ in (6), we show (5) holds. Using the constant stepsize λ as defined in (7), we have that relation (5) implies

$$\mathbb{E}_{\xi_{[k-1]}}[\phi(\bar{x}_k)] - \phi(x_*) \le \frac{d_0^2}{2\lambda k} + 8\lambda M^2 + \lambda\sigma^2 \le \frac{d_0^2}{2\lambda k} + \frac{\varepsilon}{2}.$$

The last conclusion of the theorem follows from the above inequality and (7). □

Corollary 1. *Assume $\lambda_k = \lambda$ is as in (7), then the complexity to find $\bar{x}_k$ such that*

$$\mathbb{P}(\phi(\bar{x}_k) - \phi_* \le \varepsilon) \ge 1 - p,$$

where $p \in (0, 1)$, is

$$\mathcal{O}\left(\max\left\{\frac{Ld_0^2}{\varepsilon p}, \frac{(M^2 + \sigma^2)d_0^2}{\varepsilon^2 p^2}\right\}\right). \tag{8}$$

Proof. If we have

$$\mathbb{E}[\phi(\bar{x}_k)] - \phi_* \leq p\varepsilon, \tag{9}$$

then it follows from the Markov's inequality that

$$\mathbb{P}(\phi(\bar{x}_k) - \phi_* \geq \varepsilon) \leq \frac{\mathbb{E}[\phi(\bar{x}_k)] - \phi_*}{\varepsilon} \leq p.$$

Hence, using Theorem 1, we obtain the complexity to find $\bar{x}_k$ such that (9) holds is (8). Therefore, the corollary follows. □

2.2 High probability result

It is possible, however, to obtain much finer bounds on deviation probabilities when imposing more restrictive assumptions on the distribution of $s(x, \xi)$. Specifically, assume the following "light-tail" condition.

Assumption 1. *For any $x \in \operatorname{dom} h$, we have*

$$\mathbb{E}\left[\exp\left(\|s(x,\xi) - \nabla f(x)\|^2/\sigma^2\right)\right] \leq \exp(1).$$

It can be seen that Assumption 1 implies (A3). Indeed, if a random variable X satisfies $\mathbb{E}[\exp(X/a)] \leq \exp(1)$ for some $a > 0$, then by Jensen's inequality

$$\exp(\mathbb{E}[X/a]) \leq \mathbb{E}[\exp(X/a)] \leq \exp(1),$$

and thus $\mathbb{E}[X] \leq a$. Of course, Assumption 1 holds if $\|s(x,\xi) - \nabla f(x)\|^2 \leq \sigma^2$ for all $x \in \operatorname{dom} h$ and almost every $\xi \in \Xi$.

Assumption 1 is sometimes called the sub-Gaussian assumption. Many different random variables, such as Gaussian, uniform, and any random variables with a bounded support, will satisfy this assumption.

The following result is well-known for the martingale-difference sequence.

Lemma 2. *Let $\xi_{[k]} \equiv \{\xi_1, \xi_2, \ldots, \xi_k\}$ be a sequence of i.i.d. random variables, and $\zeta_k = \zeta_k\left(\xi_{[k]}\right)$ be deterministic Borel functions of $\xi_{[k]}$ such that $\mathbb{E}\left[\zeta_k \mid \xi_{[k-1]}\right] = 0$ a.s. and $\mathbb{E}\left[\exp\left(\zeta_k^2/\sigma_k^2\right) \mid \xi_{[k-1]}\right] \leq \exp(1)$ a.s., where $\sigma_k > 0$ are deterministic. Then for any $\gamma \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^k \zeta_i > \gamma\sqrt{\sum_{i=1}^k \sigma_i^2}\right) \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Proof. Denote $\bar{\zeta}_k = \zeta_k/\sigma_k$. Then, we have

$$\mathbb{E}[\bar{\zeta}_k \mid \xi_{[k-1]}] = 0, \quad \mathbb{E}[\exp(\bar{\zeta}_k) \mid \xi_{[k-1]}] \leq \exp(1). \tag{10}$$

Also note that $\exp(x) \leq x + \exp(9x^2/16)$ for all $x \in \mathbb{R}$. Using the above relation with $x = \alpha\bar{\zeta}_k$ for $\alpha \in [0, 4/3]$ and (10), we have

$$\mathbb{E}[\exp(\alpha\bar{\zeta}_k) \mid \xi_{[k-1]}] \leq \mathbb{E}[\exp(9\alpha^2\bar{\zeta}_k^2/16) \mid \xi_{[k-1]}] \leq \exp\left(\frac{9\alpha^2}{16}\right). \tag{11}$$

It follows from the fact that $\alpha x \leq \frac{3}{8}\alpha^2 + \frac{2}{3}x^2$ that

$$\mathbb{E}[\exp(\alpha\bar{\zeta}_k) \mid \xi_{[k-1]}] \leq \exp\left(\frac{3\alpha^2}{8}\right)\mathbb{E}[\exp(2\bar{\zeta}_k^2/3) \mid \xi_{[k-1]}] \leq \exp\left(\frac{3\alpha^2}{8} + \frac{2}{3}\right),$$

and hence that for $\alpha \geq 4/3$,

$$\mathbb{E}[\exp(\alpha\bar{\zeta}_k) \mid \xi_{[k-1]}] \leq \exp\left(\frac{3\alpha^2}{4}\right). \tag{12}$$

Combining (11) and (12), we have for every $\alpha \geq 0$,

$$\mathbb{E}[\exp(\alpha\bar{\zeta}_k) \mid \xi_{[k-1]}] \leq \exp\left(\frac{3\alpha^2}{4}\right),$$

or euivalently every $t \geq 0$,

$$\mathbb{E}[\exp(t\zeta_k) \mid \xi_{[k-1]}] \leq \exp\left(\frac{3t^2\sigma_k^2}{4}\right).$$

Since ζ_k is a deterministic function of $\xi_{[k]}$, we have the recurrence

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t\sum_{i=1}^{k}\zeta_i\right)\right] &= \mathbb{E}\left[\exp\left(t\sum_{i=1}^{k-1}\zeta_i\right)\mathbb{E}[\exp(t\zeta_k) \mid \xi_{[k-1]}]\right] \\
&\leq \exp\left(\frac{3t^2\sigma_k^2}{4}\right)\mathbb{E}\left[\exp\left(t\sum_{i=1}^{k-1}\zeta_i\right)\right].
\end{aligned}$$

Hence, we have for every $t \geq 0$,

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^{k}\zeta_i\right)\right] \leq \exp\left(\frac{3t^2\sum_{i=1}^{k}\sigma_i^2}{4}\right).$$

Applying the Chebyshev's inequality, we have for $\gamma > 0$ and every $t \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{k}\zeta_i \geq \gamma\sqrt{\sum_{i=1}^{k}\sigma_i^2}\right) &\leq \exp\left(-t\gamma\sqrt{\sum_{i=1}^{k}\sigma_i^2}\right)\mathbb{E}\left[\exp\left(t\sum_{i=1}^{k}\zeta_i\right)\right] \\
&\leq \exp\left(-t\gamma\sqrt{\sum_{i=1}^{k}\sigma_i^2}\right)\exp\left(\frac{3t^2\sum_{i=1}^{k}\sigma_i^2}{4}\right).
\end{aligned}$$

Since the above ineqaulity holds for every $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{k}\zeta_i \geq \gamma\sqrt{\sum_{i=1}^{k}\sigma_i^2}\right) \leq \inf_{t\geq 0}\exp\left(-t\gamma\sqrt{\sum_{i=1}^{k}\sigma_i^2}\right)\exp\left(\frac{3t^2\sum_{i=1}^{k}\sigma_i^2}{4}\right)$$
$$= \exp\left(-\frac{\gamma^2}{3}\right).$$

□

Theorem 2. *Assume* $\operatorname{dom} h$ *has finite diameter* D. *Then, for every* $\gamma \geq 0$, *the average point* $\bar{x}_k$ *as in* (6) *satisfies*

$$\mathbb{P}\left(\phi(\bar{x}_k) - \phi_* \geq \left[2\sum_{i=0}^{k-1}\lambda_i\right]^{-1}\left[d_0^2 + \sum_{i=0}^{k-1}\frac{4\lambda_i^2M^2}{1-2\lambda_iL} + 2\gamma D\sigma\sqrt{\sum_{i=0}^{k-1}\lambda_i^2} + 2(1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right]\right)$$
$$\leq \exp\left(-\frac{\gamma^2}{3}\right) + \exp(-\gamma). \tag{13}$$

Proof. Let $\zeta_k = \lambda_k\langle s(x_k;\xi_k) - f'(x_k), x_* - x_k\rangle$ and $\Delta_k = \|s(x_k;\xi_k) - f'(x_k)\|$. Then, (4) becomes

$$\lambda_k\left(\phi(x_{k+1}) - \phi(x_*)\right) \leq \frac{1}{2}\|x_k - x_*\|^2 - \frac{1}{2}\|x_{k+1} - x_*\|^2 + \frac{4\lambda_k^2M^2}{1-2\lambda_kL} + \zeta_k + \lambda_k^2\Delta_k^2.$$

Summing the above inequality over iterations gives

$$\phi(\bar{x}_k) - \phi(x_*) \leq \frac{d_0^2 + \sum_{i=0}^{k-1}\frac{4\lambda_i^2M^2}{1-2\lambda_iL} + \sum_{i=0}^{k-1}2\zeta_i + \sum_{i=0}^{k-1}2\lambda_i^2\Delta_i^2}{2\sum_{i=0}^{k-1}\lambda_i}. \tag{14}$$

Clearly, it follows (A2) that $\mathbb{E}\left[\zeta_k \mid \xi_{[k-1]}\right] = 0$, i.e., $\{\zeta_k\}$ is a martingale-difference sequence. Moreover, it follows from the Cauchy-Schwarz inequality, the boundedness of $\operatorname{dom} h$, and Assumption 1 that

$$\mathbb{E}\left[\exp\left\{\zeta_k^2/(\lambda_k D\sigma)^2\right\} \mid \xi_{[k-1]}\right] \leq \mathbb{E}\left[\exp\left\{(\lambda_k D\Delta_k)^2/(\lambda_k D\sigma)^2\right\} \mid \xi_{[k-1]}\right] \leq \exp(1).$$

Using the previous two observations and Lemma 2, we have for every $\gamma \geq 0$,

$$\mathbb{P}\left(\sum_{i=0}^{k-1}\zeta_i > \gamma D\sigma\sqrt{\sum_{i=0}^{k-1}\lambda_i^2}\right) \leq \exp\left(-\frac{\gamma^2}{3}\right). \tag{15}$$

It follows from the convexity of $\exp(\cdot)$ that

$$\exp\left\{\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2 \Big/ \left(\sigma^2\sum_{i=1}^{k-1}\lambda_i^2\right)\right\} \leq \sum_{i=1}^{k-1}\frac{\lambda_i^2}{\sum_{i=1}^{k-1}\lambda_i^2}\exp(\Delta_i^2/\sigma^2).$$

Taking expectation of the above inequality and using Assumption 1, i.e., $\mathbb{E}[\exp(\Delta_i^2/\sigma^2)] \leq \exp(1)$, we obtain

$$\mathbb{E}\left[\exp\left\{\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2 \Big/ \left(\sigma^2\sum_{i=1}^{k-1}\lambda_i^2\right)\right\}\right] \leq \exp(1).$$

This inequality and the Markov's inequality imply that for every $\gamma \geq 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2 \geq (1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right) &= \mathbb{P}\left(\exp\left\{\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2\right\} \geq \exp\left\{(1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right\}\right)\\ &\leq \mathbb{E}\left[\exp\left(\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2\right)\right] \Big/ \exp\left((1+\gamma)\sigma^2\sum_{i=1}^{k-1}\lambda_i^2\right)\\ &\leq \exp(-\gamma).\end{aligned} \tag{16}$$

Now, we are ready to summarize the results. It follows from (14) that

$$\begin{aligned}&\mathbb{P}\left(\phi(\bar{x}_k) - \phi_* \geq \left[2\sum_{i=0}^{k-1}\lambda_i\right]^{-1}\left[d_0^2 + \sum_{i=0}^{k-1}\frac{4\lambda_i^2M^2}{1-2\lambda_iL} + 2\gamma D\sigma\sqrt{\sum_{i=0}^{k-1}\lambda_i^2} + 2(1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right]\right)\\ \leq&\mathbb{P}\left(\sum_{i=0}^{k-1}\zeta_i + \sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2 \geq \gamma D\sigma\sqrt{\sum_{i=0}^{k-1}\lambda_i^2} + (1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right)\\ \leq&\mathbb{P}\left(\sum_{i=0}^{k-1}\zeta_i > \gamma D\sigma\sqrt{\sum_{i=0}^{k-1}\lambda_i^2}\right) + \mathbb{P}\left(\sum_{i=0}^{k-1}\lambda_i^2\Delta_i^2 \geq (1+\gamma)\sigma^2\sum_{i=0}^{k-1}\lambda_i^2\right),\end{aligned}$$

where in the second inequality we use the fact that

$$\mathbb{P}(X+Y \geq a+b) \leq \mathbb{P}(\{X \geq a\} \cup \{Y \geq b\}) \leq \mathbb{P}(X \geq a) + \mathbb{P}(Y \geq b).$$

It immediately follows from (15) and (16) that (13) holds. □

Corollary 2. *Assume* dom h *has finite diameter* D *and*

$$\lambda_k = \lambda = \min\left\{\frac{\varepsilon}{2[4M^2 + (1-\log p)\sigma^2]}, \frac{1}{4L}\right\},$$

then the complexity to find $\bar{x}_k$ *such that*

$$\mathbb{P}(\phi(\bar{x}_k) - \phi_* \leq \varepsilon) \geq 1-p,$$

where $p \in (0,1)$, *is*

$$\mathcal{O}\left(\max\left\{\frac{Ld_0^2}{\varepsilon}, \frac{[M^2 + \sigma^2(1+\log\frac{1}{p})]d_0^2}{\varepsilon^2}, \frac{D^2\sigma^2}{\varepsilon^2}\log\frac{1}{p}\right\}\right). \tag{17}$$

Proof. Let $\gamma = \Theta(\log 1/p)$. In view of Theorem 2, it suffices to derive the bound on k for

$$\frac{d_0^2}{2\lambda k} + 4\lambda M^2 + \frac{D\sigma}{\sqrt{k}} \log\frac{1}{p} + \left(1 + \log\frac{1}{p}\right)\sigma^2\lambda \le \varepsilon.$$

Using the choice of λ, it boils down to deriving the bound on k for

$$\frac{d_0^2}{2\lambda k} + \frac{D\sigma}{\sqrt{k}} \log\frac{1}{p} \le \frac{\varepsilon}{2}.$$

Hence, we prove (17) is the complexity bound to find $\bar{x}_k$ such that it is an ε-solution with probability at least $1 - p$. $\square$