Introduction to Zeroth-Order Optimization

Yujie Tang

1 Review of Gradient Descent

Consider the following unconstrained optimization problem

$$\min_{x\in\mathbb{R}^p} \ f(x) \tag{1}$$

where $f:\mathbb{R}^p\to\mathbb{R}$ is continuously differentiable. The gradient descent iteration for minimizing $f(x)$ over $x\in\mathbb{R}^p$ is given by

$$x_{k+1} = x_k - \alpha\nabla f(x_k), \tag{GD}$$

where $\alpha>0$ is the step size. The following theorem establishes the convergence of gradient descent for smooth and convex objective functions.

Theorem 1. *Suppose $f:\mathbb{R}^p\to\mathbb{R}$ is convex and L-smooth, and has a minimizer $x^*\in\mathbb{R}^p$.*

1. *By choosing $\alpha = 1/L$, the gradient descent iteration* (GD) *achieves*

$$f(x_k) - f(x^*) \leq \frac{L\|x_0-x^*\|^2}{2(k+1)}.$$

2. *If f is also m-strongly convex, then by choosing $\alpha = 2/(L+m)$, the gradient descent iteration* (GD) *achieves*

$$\|x_k-x^*\| \leq \left(\frac{L-m}{L+m}\right)^k \|x_0-x^*\|, \quad f(x_k)-f(x^*) \leq \frac{L}{2}\left(\frac{L-m}{L+m}\right)^{2k}\|x_0-x^*\|^2.$$

Corollary 1. *Suppose $f:\mathbb{R}^p\to\mathbb{R}$ is convex and L-smooth, and has a minimizer $x^*\in\mathbb{R}^p$. Let $\epsilon>0$ be arbitrary.*

1. *The number of gradient descent iterations needed to achieve $f(x_k)-f(x^*)\leq\epsilon$ can be bounded by*

$$k = O\!\left(\frac{1}{\epsilon}\right).$$

2. *If f is also m-strongly convex, then the number of gradient descent iterations needed to achieve $f(x_k)-f(x^*)\leq\epsilon$ can be bounded by*

$$k = O\!\left(\ln\frac{1}{\epsilon}\right).$$

2 Zeroth-Order Gradient Estimation

Now suppose we don't have access to the gradients of the function f. Instead, there is a zeroth-order oracle that can accept an arbitrary $x \in \mathbb{R}^p$ and output the corresponding value $f(x)$, and we can only employ this zeroth-order oracle finitely many times for optimizing f. In this lecture, we introduce a class of methods based on gradient estimation using zeroth-order information.

We start with the following single-point zeroth-order gradient estimator:

$$\mathsf{G}_f(x; r, z) = \frac{p}{r} f(x + rz)\, z, \qquad z \sim \mathcal{Z}. \tag{2}$$

Here $r > 0$ is a positive parameter called the *smoothing radius*; z is a p-dimensional random vector following the probability distribution $\mathcal{Z}$, and we will just call it the *random perturbation*. Usually, the $\mathcal{Z}$ is chosen to be one of the following:

1. The Gaussian distribution $\mathcal{N}(0, p^{-1}I)$.
2. The uniform distribution on the unit sphere $\mathbb{S}_{p-1} := \{x \in \mathbb{R}^p : \|x\| = 1\}$, which we denote by $\mathrm{Unif}(\mathbb{S}_{p-1})$.

The following lemma characterizes the expectation of the single-point gradient estimator (2).

Lemma 1. *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is L-smooth.*

1. Let $\mathcal{Z}$ be $\mathcal{N}(0, p^{-1}I)$. Then

$$\mathbb{E}_{z\sim\mathcal{Z}}[\mathsf{G}_f(x; r, z)] = \nabla f_r(x),$$

where $f_r : \mathbb{R}^p \to \mathbb{R}$ is given by

$$f_r(x) := \mathbb{E}_{y\sim\mathcal{Y}}[f(x + ry)],$$

and $\mathcal{Y}$ is the Gaussian distribution $\mathcal{N}(0, p^{-1}I)$.

2. let $\mathcal{Z}$ be $\mathrm{Unif}(\mathbb{S}_{p-1})$. Then

$$\mathbb{E}_{z\sim\mathcal{Z}}[\mathsf{G}_f(x; r, z)] = \nabla f_r(x),$$

where $f_r : \mathbb{R}^p \to \mathbb{R}$ is given by

$$f_r(x) := \mathbb{E}_{y\sim\mathcal{Y}}[f(x + ry)],$$

and $\mathcal{Y}$ is the uniform distribution on the unit ball $\mathbb{B}_p := \{x \in \mathbb{R}^p : \|x\| \le 1\}$.

Lemma 1 shows that the expectation of $\mathsf{G}_f(x; r, z)$ gives the gradient of a *smooth version* of f. The following lemma provides further properties of f_r and ∇f_r.

Lemma 2. *Suppose f is convex and L-smooth. Let $\mathcal{Z}$ be either $\mathcal{N}(0, p^{-1}I)$ or $\mathrm{Unif}(\mathbb{S}_{p-1})$, and let f_r denote the corresponding smooth version of f. Then f_r is convex, L-smooth, and satisfies*

$$f(x) \le f_r(x) \le f(x) + \frac{Lr^2}{2},$$

and

$$\|\nabla f_r(x) - \nabla f(x)\| \le Lr.$$

Proof. The convexity of f_r follows by noting that

$$\begin{aligned} f_r(\theta x_1 + (1-\theta)x_2) &= \mathbb{E}_{y\sim\mathcal{Y}}[f(\theta x_1 + (1-\theta)x_2 + ry)] \\ &= \mathbb{E}_{y\sim\mathcal{Y}}[f(\theta(x_1 + ry) + (1-\theta)(x_2 + ry))] \\ &\leq \mathbb{E}_{y\sim\mathcal{Y}}[\theta f(x_1 + ry) + (1-\theta) f(x_2 + ry)] = \theta f_r(x_1) + (1-\theta) f_r(x_2) \end{aligned}$$

for any $\theta \in [0,1]$ and any x_1, x_2.

To show the L-smoothness of f_r, let $x_1, x_2 \in \mathbb{R}^p$ be arbitrary, and we have

$$\begin{aligned} \|\nabla f_r(x_1) - \nabla f_r(x_2)\| &= \|\nabla\,\mathbb{E}_{y\sim\mathcal{Y}}[f(x_1 + ry)] - \nabla\,\mathbb{E}_{y\sim\mathcal{Y}}[f(x_2 + ry)]\| \\ &= \|\mathbb{E}_{y\sim\mathcal{Y}}[\nabla f(x_1 + ry) - \nabla f(x_2 + ry)]\| \\ &\leq \mathbb{E}_{y\sim\mathcal{Y}}[\|\nabla f(x_1 + ry) - \nabla f(x_2 + ry)\|] \\ &\leq \mathbb{E}_{y\sim\mathcal{Y}}[L\|x_1 - x_2\|] = L\|x_1 - x_2\|, \end{aligned}$$

where in the second step we interchanged differentiation with expectation (which can be justified by the dominance convergence theorem), and in the fourth step we used the L-smoothness of f.

Now, by the convexity and smoothness of f, we have

$$f(x) + \langle \nabla f(x), ry\rangle \leq f(x + ry) \leq f(x) + \langle \nabla f(x), ry\rangle + \frac{L}{2}\|ry\|^2.$$

Now we take the expectation with respect to $y \sim \mathcal{Y}$. We have $\mathbb{E}_{y\sim\mathcal{Y}}[\langle \nabla f(x), ry\rangle] = 0$ since $\mathcal{Y}$ is isotropic, and therefore

$$f(x) \leq \mathbb{E}_{y\sim\mathcal{Y}}[f(x + ry)] \leq f(x) + \frac{Lr^2}{2}\mathbb{E}_{y\sim\mathcal{Y}}\left[\|y\|^2\right],$$

which gives the first inequality.

Now regarding ∇f_r, we have

$$\begin{aligned} \|\nabla f_r(x) - \nabla f(x)\| &= \|\nabla_x\,\mathbb{E}_{y\sim\mathcal{Y}}[f(x + ry) - f(x)]\| \\ &= \|\mathbb{E}_{y\sim\mathcal{Y}}[\nabla_x f(x + ry) - \nabla_x f(x)]\| \\ &\leq \mathbb{E}_{y\sim\mathcal{Y}}[\|\nabla f(x + ry) - \nabla f(x)\|] \\ &\leq \mathbb{E}_{y\sim\mathcal{Y}}[L\|ry\|] \leq Lr, \end{aligned}$$

where in the second step we interchanged differentiation with expectation (which can be justified by the dominance convergence theorem), and in the fourth step we employ the L-smoothness of f. □

Lemma 2 bounds the differences $f_r - f$ and $\nabla f_r - \nabla f$, and we can see that they both go to zero when $r \to 0$. Consequently, we can view $\mathsf{G}_f(x; r, z)$ as a stochastic gradient of f with a bias that can be controlled by the smoothing radius r.

3 Two-Point Gradient Estimators

The single-point gradient estimator (2) provides a stochastic gradient with a nonzero but controllable bias. However, its variance (or second-moment) is roughly on the order of r^{-2}, which can be large and can slow down convergence. In this section, we study a popular variant of the single-point gradient estimator, which we call the two-point zeroth-order gradient estimators, that employ two function values for reducing the variance.

There are two versions of two-point gradient estimators, which are

$$\mathsf{G}_f^{(2)}(x; r, z) = \frac{p}{r}(f(x + rz) - f(x))\, z$$

and

$$\tilde{\mathsf{G}}_f^{(2)}(x; r, z) = \frac{p}{2r}(f(x + rz) - f(x - rz))\, z,$$

where $z \sim \mathcal{Z}$ is again a random perturbation and $\mathcal{Z}$ is usually either $\mathrm{Unif}(\mathbb{S}_{p-1})$ or $\mathcal{N}(0, p^{-1}I)$. Since $\mathcal{Z}$ is isotropic, we can see that both $\mathsf{G}_f^{(2)}(x; r, z)$ and $\tilde{\mathsf{G}}_f^{(2)}(x; r, z)$ have the same expectation as the single-point one, i.e.,

$$\mathbb{E}_{z\sim\mathcal{Z}}\left[\mathsf{G}_f^{(2)}(x; r, z)\right] = \mathbb{E}_{z\sim\mathcal{Z}}\left[\tilde{\mathsf{G}}_f^{(2)}(x; r, z)\right] = \nabla f_r(x).$$

On the other hand, the following lemma shows that their second-moments have better dependencies on the smoothing radius r.

Lemma 3. *Suppose f is L-smooth, and let $\mathcal{Z}$ be either $\mathcal{N}(0, p^{-1}I)$ or $\mathrm{Unif}(\mathbb{S}_{p-1})$. Then*

$$\mathbb{E}_{z\sim\mathcal{Z}}\left[\left\|\mathsf{G}_f^{(2)}(x; r, z)\right\|^2\right] \le \begin{cases} 2(p+2)\|\nabla f(x)\|^2 + \dfrac{r^2L^2p^2}{2}\left(\dfrac{p+6}{p}\right)^3, & \mathcal{Z} \text{ is } \mathcal{N}(0, p^{-1}I), \\ 2p\|\nabla f(x)\|^2 + \dfrac{r^2L^2p^2}{2}, & \mathcal{Z} \text{ is } \mathrm{Unif}(\mathbb{S}_{p-1}), \end{cases}$$

and the same bound holds for $\mathbb{E}_{z\sim\mathcal{Z}}\left[\left\|\tilde{\mathsf{G}}_f^{(2)}(x; r, z)\right\|^2\right]$.

Proof. We only give a proof for $\mathsf{G}_f^{(2)}(x; r, z)$.

We have

$$\begin{aligned} \mathbb{E}_z\left[\left\|\mathsf{G}_f^{(2)}(x; r, z)\right\|^2\right] &= \frac{p^2}{r^2}\,\mathbb{E}_z\left[|f(x + rz) - f(x)|^2 \cdot \|z\|^2\right] \\ \le \frac{p^2}{r^2}\,\mathbb{E}_z&\left[\left(2\,|f(x + rz) - f(x) - \langle \nabla f(x), rz\rangle|^2 + 2|\langle \nabla f(x), rz\rangle|^2\right)\|z\|^2\right] \\ = \frac{2p^2}{r^2}\,\mathbb{E}_z&\left[|f(x + rz) - f(x) - \langle \nabla f(x), rz\rangle|^2\,\|z\|^2\right] + 2p^2\mathbb{E}_z\left[|\langle \nabla f(x), z\rangle|^2\|z\|^2\right] \end{aligned} \tag{3}$$

First we consider the second term in (3). Note that

$$\mathbb{E}_z\left[|\langle \nabla f(x), z\rangle|^2 \cdot \|z\|^2\right] = (\nabla f(x))^\top \mathbb{E}_z\left[\|z\|^2 zz^\top\right] \nabla f(x).$$

If $\mathcal{Z}$ is the Gaussian distribution $\mathcal{N}(0, p^{-1}I)$, then

$$\mathbb{E}_z\left[\|z\|^2 z_i z_j\right] = \sum_{k=1}^{p}\mathbb{E}_z\left[z_k^2 z_i z_j\right] = \begin{cases} \dfrac{p+2}{p^2}, & i = j, \\ 0, & i \ne j, \end{cases}$$

where we used $\mathbb{E}_z[z_i^4] = 3/p^2$, and therefore

$$\mathbb{E}_z\left[\|z\|^2 zz^\top\right] = \frac{p+2}{p^2} I.$$

If $\mathcal{Z}$ is $\mathrm{Unif}(\mathbb{S}_{p-1})$, then

$$\mathbb{E}_z\left[\|z\|^2 zz^\top\right] = \mathbb{E}_z\left[zz^\top\right] = \frac{1}{p} I,$$

where we used $\mathbb{E}_z[z_i z_j] = 0$ for $i = j$ by the symmetry of $\mathcal{Z}$. Therefore

$$2p^2 \mathbb{E}_z\left[|\langle \nabla f(x), z\rangle|^2 \cdot \|z\|^2\right] = \begin{cases} 2(p+2)\|\nabla f(x)\|^2, & \mathcal{Z} \text{ is } \mathcal{N}(0, p^{-1}I), \\ 2p\|\nabla f(x)\|^2, & \mathcal{Z} \text{ is } \mathrm{Unif}(\mathbb{S}_{p-1}). \end{cases}$$

Next we bound the first term. By Newton-Leibniz theorem,

$$f(x+rz) - f(x) = \int_0^r \langle \nabla f(x+tz), z\rangle \, dt,$$

and thus

$$\begin{aligned} |f(x+rz) - f(x) - \langle \nabla f(x), rz\rangle| &= \left| \int_0^r \langle \nabla f(x+tz) - \nabla f(x), z\rangle \, dt \right| \\ &\le \int_0^r \|\nabla f(x+tz) - \nabla f(x)\| \|z\| \, dt \\ &\le \int_0^r Lt\|z\|^2 \, dt = \frac{Lr^2}{2}\|z\|^2. \end{aligned}$$

We then get

$$\begin{aligned} &\frac{2p^2}{r^2} \mathbb{E}_z\left[|f(x+rz) - f(x) - \langle f(x), rz\rangle|^2 \|z\|^2\right] \\ &\le \frac{2p^2}{r^2} \mathbb{E}_z\left[\frac{L^2 r^4}{4}\|z\|^6\right] \le \begin{cases} \left(\dfrac{p+6}{p}\right)^3 \dfrac{r^2 L^2 p^2}{2}, & \mathcal{Z} \text{ is } \mathcal{N}(0, p^{-1}I), \\ \dfrac{r^2 L^2 p^2}{2}, & \mathcal{Z} \text{ is } \mathrm{Unif}(\mathbb{S}_{p-1}), \end{cases} \end{aligned}$$

where we used $\mathbb{E}_z[\|z\|^6] \le (p+6)^3/p^3$ for $z \sim \mathcal{N}(0, p^{-1}I)$. □

Lemma 3 shows that the second-moment of either of the two-point gradient estimators does not blow up as $r \to 0$,[1] and thus achieves much smaller variance compared to the single-point gradient estimator for small r.

[1] For practical numerical computation, however, r cannot be arbitrarily small due to machine precision.

4 Convergence Analysis for Zeroth-Order Optimization

We now turn our focus to convergence analysis of zeroth-order optimization method, and study the following iteration as an example:

$$x_{k+1} = x_k - \alpha\, \mathsf{G}_f^{(2)}(x_k; r_k, z_k), \tag{4}$$

i.e., we plug the two-point gradient estimator $\mathsf{G}_f^{(2)}$ into the stochastic gradient descent iteration. Here each z_k is independently drawn from the distribution $\mathcal{Z}$, and r_k is a positive sequence of smoothing radii that vary with k. We let $\mathcal{Z}$ be $\mathrm{Unif}(\mathbb{S}_{p-1})$ for simplicity. We assume that f is convex and L-smooth and has a minimizer $x \in \mathbb{R}^p$.

Let $\mathcal{F}_k$ denote the filtration generated by $(x_1, \ldots, x_k)$. Our convergence analysis starts by expanding $\|x_{k+1} - x^*\|^2$:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha\, \mathsf{G}_f^{(2)}(x_k; r_k, z_k)\|^2 \\
&= \|x_k - x^*\|^2 - 2\alpha \left\langle x_k - x^*, \mathsf{G}_f^{(2)}(x_k; r_k, z_k) \right\rangle + \alpha^2 \left\| \mathsf{G}_f^{(2)}(x_k; r_k, z_k) \right\|^2.
\end{aligned}$$

By taking the expectation conditioned on $\mathcal{F}_k$ and using Lemma 1 and Lemma 3, we get

$$\mathbb{E}\Big[\Big\langle x_k - x^*, \mathsf{G}_f^{(2)}(x_k; r_k, z_k) \Big\rangle \Big| \mathcal{F}_k \Big] = \langle x_k - x^*, \nabla f_{r_k}(x_k) \rangle,$$

$$\mathbb{E}\left[\left\| \mathsf{G}_f^{(2)}(x_k; r_k, z_k) \right\|^2 \middle| \mathcal{F}_k \right] \le 2p \|\nabla f(x_k)\|^2 + \frac{r_k^2 L^2 p^2}{2},$$

and consequently

$$\mathbb{E}\big[\|x_{k+1} - x^*\|^2 \big| \mathcal{F}_k \big] \le \|x_k - x^*\|^2 - 2\alpha \langle x_k - x^*, \nabla f_{r_k}(x_k) \rangle + 2\alpha^2 p \|\nabla f(x_k)\|^2 + \frac{\alpha^2 r_k^2 L^2 p^2}{2}.$$

Since f_{r_k} is convex, we see that

$$f_{r_k}(x^*) - f_{r_k}(x_k) \ge \langle \nabla f_r(x_k), x^* - x_k \rangle,$$

and by Lemma 2, we further have

$$-\langle \nabla f_{r_k}(x_k), x_k - x^* \rangle \le f_{r_k}(x^*) - f_{r_k}(x_k) \le f(x^*) - f(x_k) + \frac{L r_k^2}{2}.$$

Moreover, since f is L-smooth and $\nabla f(x^*) = 0$, we have

$$\|\nabla f(x_k)\|^2 = \|\nabla f(x_k) - \nabla f(x^*)\|^2 \le 2L(f(x_k) - f(x^*)).$$

Summarizing these results, we get

$$\mathbb{E}\big[\|x_{k+1} - x^*\|^2 \big| \mathcal{F}_k \big] \le \|x_k - x^*\|^2 - 2\alpha(1 - 2\alpha p L)(f(x_k) - f(x^*)) + \alpha L r_k^2 + \frac{\alpha^2 r_k^2 L^2 p^2}{2},$$

and by taking the total expectation, we can get

$$2\alpha(1 - 2\alpha p L)\mathbb{E}[f(x_k) - f(x^*)] \le \mathbb{E}\big[\|x_k - x^*\|^2\big] - \mathbb{E}\big[\|x_{k+1} - x^*\|^2\big] + \alpha L r_k^2 + \frac{\alpha^2 r_k^2 L^2 p^2}{2}.$$

Now we take the telescoping sum and get

$$2\alpha(1-2\alpha pL)\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f(x^*)] \le \|x_0-x^*\|^2 + \alpha L\left(1+\frac{\alpha Lp^2}{2}\right)\sum_{k=0}^{K} r_k^2.$$

By taking $\alpha = c/(2pL)$ for some $c \in (0,1)$, we get

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f(x^*)] \le \frac{pL\|x_0-x^*\|^2}{c(1-c)(K+1)} + \frac{L}{2(1-c)}\left(1+\frac{cp}{4}\right)\frac{\sum_{k=0}^{K} r_k^2}{K+1},$$

which further implies

$$\mathbb{E}\left[\min_{0\le k\le K} f(x_k)-f(x^*)\right] \le \frac{pL\|x_0-x^*\|^2}{c(1-c)(K+1)} + \frac{L}{2(1-c)}\left(1+\frac{cp}{4}\right)\frac{\sum_{k=0}^{K} r_k^2}{K+1}.$$

The following theorem summarizes the convergence analysis of the iteration (4) for the smooth and convex setting.

Theorem 2. *Suppose f is convex and L-smooth, and has a minimizer $x \in \mathbb{R}^p$. Let $\alpha = c/(pL)$ for some $c \in (0,1)$, and let r_k be a positive sequence of smoothing radii such that $\sum_{k=0}^{K} r_k^2 = R^2 < +\infty$. Then the zeroth-order optimization iteration (4) achieves*

$$\mathbb{E}\left[\min_{0\le k\le K} f(x_k)-f(x^*)\right] \le \frac{p}{K+1}\left(\frac{L\|x_0-x^*\|^2}{c(1-c)} + \frac{R^2L(c+4/p)}{8(1-c)}\right).$$

Corollary 2. *Let $\epsilon > 0$ be arbitrary. Then, under the conditions of Theorem (2), the number of zeroth-order queries needed to achieve*

$$\mathbb{E}\left[\min_{0\le k\le K} f(x_k)-f(x^*)\right] \le \epsilon$$

is bounded by

$$2(K+1) = O\Big(\frac{p}{\epsilon}\Big).$$

Remark 1. For smooth constrained convex optimization, the best convergence rate established so far seems to be $O(\sqrt{p/K})$ (or $O(p/\epsilon^2)$ in terms of iteration complexity), which is worse than the unconstrained case. This is different from first-order methods where projected gradient descent can still achieve $O(1/K)$ convergence rate for smooth constrained convex objectives.

4.1 Convergence Analysis for Smooth and Strongly Convex f

Theorem 3. *Suppose f is m-strongly convex and L-smooth, and has a minimizer $x^* \in \mathbb{R}^p$. Let $\alpha = c/(pL)$ for some $c \in (0,1)$. Then the zeroth-order optimization iteration (4) achieves*

$$\mathbb{E}\big[\|x_k-x^*\|^2\big] \le \rho^k\|x_0-x^*\|^2 + \frac{c(c+4/p)}{8}\sum_{\tau=0}^{k-1}\rho^\tau r_{k-1-\tau}^2,$$

where

$$\rho = 1-\frac{c(1-c)m}{2pL}.$$

Proof. Much of the derivation for the smooth and convex setting can be applied here, and we have

$$\mathbb{E}\big[\,\|x_{k+1}-x^*\|^2\big|\,\mathcal{F}_k\big] \le \|x_k-x^*\|^2 - 2\alpha(1-2\alpha pL)(f(x_k)-f(x^*)) + \alpha L r_k^2 + \frac{\alpha^2 r_k^2 L^2 p^2}{2}.$$

Now since f is m-strongly convex, we have

$$f(x_k)-f(x^*) \ge \langle \nabla f(x^*), x_k - x^*\rangle + \frac{m}{2}\|x_k-x^*\|^2 = \frac{m}{2}\|x_k-x^*\|^2.$$

Since $1-2\alpha pL > 0$, we see that

$$\mathbb{E}\big[\,\|x_{k+1}-x^*\|^2\big|\,\mathcal{F}_k\big] \le (1-\alpha m(1-2\alpha pL))\,\|x_k-x^*\|^2 + \alpha L\left(1+\frac{\alpha L p^2}{2}\right) r_k^2.$$

By plugging in $\alpha = c/(2pL)$ and taking the total expectation, we get

$$\mathbb{E}\big[\|x_{k+1}-x^*\|^2\big] \le \left(1-\frac{c(1-c)m}{2pL}\right)\mathbb{E}\big[\|x_k-x^*\|^2\big] + \frac{c(c+4/p)}{8} r_k^2.$$

The final bound can then be shown by mathematical induction. □

5 Notes and References

The main reference for this lecture is [1], which also considers nonsmooth convex optimization and nonconvex optimization problems. Some other related references on zeroth-order gradient estimation methods include:

- [2] considers constrained nonsmooth online convex optimization where only one function evaluation is available for the objective function at each time instant. The paper provides basic properties of the single-point gradient estimator.
- [3] considers unconstrained stochastic optimization where two function evaluations are available for each random sample, and employs the two-point gradient estimator $\mathsf{G}_f^{(2)}$. Convergence analysis is provided for both convex and nonconvex settings with smooth objectives.
- [4] considers constrained stochastic convex optimization where two function evaluations are available for each random sample, and combines two-point gradient estimators with the stochastic mirror descent method. It also establishes information-theoretic lower bounds on the optimal convergence rate.
- [5] considers constrained online convex optimization where two function evaluations are available for the objective function at each time instant. The paper employs $\tilde{\mathsf{G}}_f^{(2)}$ for handling convex but nonsmooth objectives, and shows that the proposed algorithm achieves the optimal convergence rate.
- [6] proposes the residual feedback method for reducing the variance of one-point gradient estimator:
  $$x_{k+1} = x_k - \alpha\cdot\frac{p}{r}(f(x_k+rz_k) - f(x_{k-1}+rz_{k-1}))z_k.$$
  This method improves on the convergence rate compared to the vanilla single-point gradient estimation method. [7] studies accelerating single-point zeroth-order methods and derive the residual feedback method from the perspective of extreme seeking control.

- [8] provides a zeroth-order stochastic coordinate descent method, in which the two-point gradient estimator $\tilde{\mathsf{G}}_f^{(2)}$ is employed, and the distribution $\mathcal{Z}$ is the uniform distribution on the standard basis $\{e_i\}_{i=1}^p$ of $\mathbb{R}^p$. The paper also considers the setting of asynchronous parallel optimization.

The paper [9] provides a recent survey of literature on zeroth-order optimization methods.

A Proof of Lemma 1

$\mathcal{Z}$ is $\mathcal{N}(0, p^{-1}I)$. In this case, we have

$$\begin{aligned} f_r(x) &= \frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} f(x+ry)\, \exp\left(-\frac{p\|y\|^2}{2}\right)\, dy \\ &= \frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} f(u) \exp\left(-\frac{p\|u-x\|^2}{2r^2}\right) \frac{1}{r^p}\, du, \end{aligned}$$

where in the second equality we substituted $u = x + ry$. We can then calculate the gradient of $f_r(x)$ by

$$\begin{aligned} \nabla f_r(x) &= \nabla_x \left(\frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} f(u) \exp\left(-\frac{p\|u-x\|^2}{2r^2}\right) \frac{1}{r^p}\, du \right) \\ &= \frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} f(u)\, \nabla_x \left(\exp\left(-\frac{p\|u-x\|^2}{2r^2}\right)\right) \frac{1}{r^p}\, du \\ &= \frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} f(u) \exp\left(-\frac{p\|u-x\|^2}{2r^2}\right) \cdot \frac{p(x-u)}{r^2} \frac{1}{r^p}\, du \\ &= \frac{1}{(2\pi/p)^{p/2}} \int_{\mathbb{R}^p} \frac{p}{r} f(x+rz) z \cdot \exp\left(-\frac{p\|z\|^2}{2}\right)\, dz = \mathbb{E}_{z\sim\mathcal{Z}}\left[\frac{p}{r} f(x+rz)z\right], \end{aligned}$$

where in the second step we interchange differentiation and integration.

$\mathcal{Z}$ is $\mathrm{Unif}(\mathbb{S}_{p-1})$. Let V_p denote the p-dimensional volume of $\mathbb{B}_p$, and let S_{p-1} denote the surface area (or $(p-1)$-dimensional volume) of $\mathbb{S}_{p-1}$. Then for any $v \in \mathbb{R}^p$, we have

$$\begin{aligned} v \cdot \nabla f_r(x) &= v \cdot \nabla_x \left(\frac{1}{V_p} \int_{\mathbb{B}_p} f(x+ry)\, dy \right) \\ &= \frac{1}{V_p} \int_{\mathbb{B}_p} v \cdot \nabla_x f(x+ry)\, dy \\ &= \frac{1}{V_p} \int_{\mathbb{B}_p} \frac{1}{r} \nabla_z \cdot (f(x+rz)v)\, dz \\ &= \frac{1}{rV_p} \int_{\mathbb{S}_{p-1}} f(x+rz) v \cdot z\, d\Sigma(z) \\ &= v \cdot \left(\frac{p}{r} \frac{1}{S_{p-1}} \int_{\mathbb{S}_{p-1}} f(x+rz) z\, d\Sigma(z) \right) = v \cdot \mathbb{E}_{z\sim\mathcal{Z}}\left[\frac{p}{r} f(x+rz)z\right]. \end{aligned}$$

Here in the fourth step, we used Gauss's divergence theorem and the fact that the unit normal vector at z on $\mathbb{S}_{p-1}$ is just z, and we use $d\Sigma(z)$ to denote the surface element of $\mathbb{S}_{p-1}$ at z; in the fifth step we used $S_{p-1} = pV_p$. By the arbitrariness of v we get the desired result.

References

[1] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[2] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

[3] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[4] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[5] O. Shamir, "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1703–1713, 2017.

[6] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.

[7] X. Chen, Y. Tang, and N. Li, "Improve single-point zeroth-order optimization using high-pass and low-pass filters from extremum seeking control," *arXiv preprint arXiv:2111.01701*, 2021.

[8] X. Lian, H. Zhang, C.-J. Hsieh, Y. Huang, and J. Liu, "A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[9] S. Liu, P.-Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero III, and P. K. Varshney, "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 43–54, 2020.