CS331-HW8-Lukang-Sun

October 29, 2021

p1.

Proof. let $u := prox_{\gamma R}(x), v := prox_{\gamma R}(y)$, then by definition we have, $0 \in \partial R(u) + \frac{1}{\gamma}(u-x), 0 \in \partial R(v) + \frac{1}{\gamma}(v-y)$, so $x-u-(y-v) \in \gamma(\partial R(u) - \partial R(v))$, by strong convexity of R, we have

$$\langle x-u-(y-v), u-v\rangle \geq \gamma\lambda||u-v||_2^2, \tag{1}$$

so

$$\begin{aligned} ||x-y||^2 &= ||x-u-(y-v)+(u-v)||^2 \\ &\geq ||u-v||^2 + 2\langle x-u-(y-v), u-v\rangle \\ &\geq (1+2\gamma\lambda)||u-v||^2. \end{aligned} \tag{2}$$

This proves the problem. □

p2.

Theorem. *Assume that f is μ-convex and R is λ-convex, g^k is unbiased (Assumption 1) and that the AC assumption (Assumption 2) is satisfied. Choose a stepsize satisfying*

$$0 < \gamma \leq \frac{1}{A}$$

Then the iterates $\left\{x^k\right\}_{k\geq 0}$ of SGD (Algorithm 3) satisfy

$$E\left[\left\|x^k - x^\star\right\|^2\right] \leq (\frac{1-\mu\gamma}{1+2\lambda\gamma})^k||x^0-x^\star||^2 + \frac{C\gamma}{\mu(1+2\lambda\gamma)}$$

Proof. let $r^k \overset{\text{def}}{=} x^k - x^\star$. Performing the exact same steps as in the case of

GD, but replacing the gradient $\nabla f\left(x^k\right)$ by the stochastic gradient g^k, we get

$$\begin{aligned}
\left\|r^{k+1}\right\|^2 &= \left\|\operatorname{prox}_{\gamma R}\left(x^k-\gamma g^k\right)-x^\star\right\|^2 \\
&= \left\|\operatorname{prox}_{\gamma R}\left(x^k-\gamma g^k\right)-\operatorname{prox}_{\gamma R}\left(x^\star-\gamma\nabla f\left(x^\star\right)\right)\right\|^2 \\
&\overset{(p1.)}{\leq} \frac{1}{1+2\gamma\lambda}\left\|\left(x^k-\gamma g^k\right)-\left(x^\star-\gamma\nabla f\left(x^\star\right)\right)\right\|^2 \\
&= \frac{1}{1+2\gamma\lambda}(\left\|x^k-x^\star-\gamma\left(g^k-\nabla f\left(x^\star\right)\right)\right\|^2) \\
&= \frac{1}{1+2\gamma\lambda}(\left\|r^k\right\|^2-2\gamma\left\langle r^k,g^k-\nabla f\left(x^\star\right)\right\rangle+\gamma^2\left\|g^k-\nabla f\left(x^\star\right)\right\|^2),
\end{aligned}$$

so

$$\begin{aligned}
\mathrm{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\leq \frac{1}{1+2\gamma\lambda}(\left\|r^k\right\|^2-2\gamma\mathrm{E}\left[\left\langle r^k,g^k-\nabla f\left(x^\star\right)\right\rangle \mid x^k\right]+\gamma^2\mathrm{E}\left[\left\|g^k-\nabla f\left(x^\star\right)\right\|^2 \mid x^k\right]) \\
&= \frac{1}{1+2\gamma\lambda}(\left\|r^k\right\|^2-2\gamma\left\langle r^k,E\left[g^k \mid x^k\right]-\nabla f\left(x^\star\right)\right\rangle+\gamma^2E\left[\left\|g^k-\nabla f\left(x^\star\right)\right\|^2 \mid x^k\right] \\
&= \frac{1}{1+2\gamma\lambda}(\left\|r^k\right\|^2-2\gamma\left\langle r^k,\nabla f\left(x^k\right)-\nabla f\left(x^\star\right)\right\rangle+\gamma^2E\left[\left\|g^k-\nabla f\left(x^\star\right)\right\|^2 \mid x^k\right]),
\end{aligned}$$

use

$$\left\langle r^k,\nabla f\left(x^k\right)-\nabla f\left(x^\star\right)\right\rangle \geq D_f\left(x^k,x^\star\right)+\frac{\mu}{2}\left\|x^k-x^\star\right\|^2,$$

lead to

$$E\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] \leq \frac{1}{1+2\gamma\lambda}\left((1-\gamma\mu)\left\|r^k\right\|^2-2\gamma D_f\left(x^k,x^\star\right)+\gamma^2E\left[\left\|g^k-\nabla f\left(x^\star\right)\right\|^2 \mid x^k\right]\right),$$

use Assumption (2), we get

$$E\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] \leq \frac{1}{1+2\lambda\gamma}(1-\gamma\mu)||r^k||^2+\frac{1}{1+2\lambda\gamma}\gamma^2C,$$

so we have

$$E\left[(1+2\lambda\gamma)^k||r^k||^2\right] \leq (1-\mu\gamma)^k||r^0||^2+\frac{\gamma C(1+2\lambda\gamma)^{k-1}}{\mu},$$

divide by $(1+2\lambda\gamma)^k$ from both sides, we have

$$E\left[||r^k||\right] \leq (\frac{1-\mu\gamma}{1+2\lambda\gamma})^k||r^0||^2+\frac{C\gamma}{\mu(1+2\lambda\gamma)}$$

□

p3. (i)

$$E\left[g^k \mid \mathcal{F}_k(y)\right] = (1-q)\nabla f(y^k) + q(\frac{1}{q}\nabla f(x^k) + (1-\frac{1}{q})\nabla f(y^k)) = \nabla f(x^k)$$

(ii)

Lemma. *Assume f is μ-convex and L-smooth, R is convex, then we have*

$$\mathrm{E}\left[\left\|g^k - \nabla f\left(x^\star\right)\right\|^2 \mid x^k, \xi^k\right] \leq 2AD_f\left(x^k, x^\star\right) + B\sigma^k,$$

$$\mathrm{E}\left[\sigma^{k+1} \mid x^k, \xi^k\right] \leq 2\tilde{A}D_f\left(x^k, x^\star\right) + \tilde{B}\sigma^k,$$

where $\sigma^k = \mathrm{E}\left[||\nabla f(y^k) - \nabla f(x^\star)||^2\right]$, $A = \frac{2L}{q}, B = 1 - q + \frac{2(1-q)^2}{q}, \tilde{A} = pL, \tilde{B} = (1-p)$.

Theorem. *Assume that f is μ-convex.For any $M > \frac{B}{1-\tilde{B}}$ choose a stepsize γ satisfying*

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + M\tilde{A}}\right\}$$

Then the iterates $\left\{x^k, \sigma^k\right\}_{k\geq 0}$ of SGD-CTRL satisfy

$$\mathrm{E}\left[V^k\right] \leq \max\left\{(1-\gamma\mu)^k, \left(\frac{B + M\tilde{B}}{M}\right)^k\right\} V^0$$

where the Lyapunov function V^k is defined by

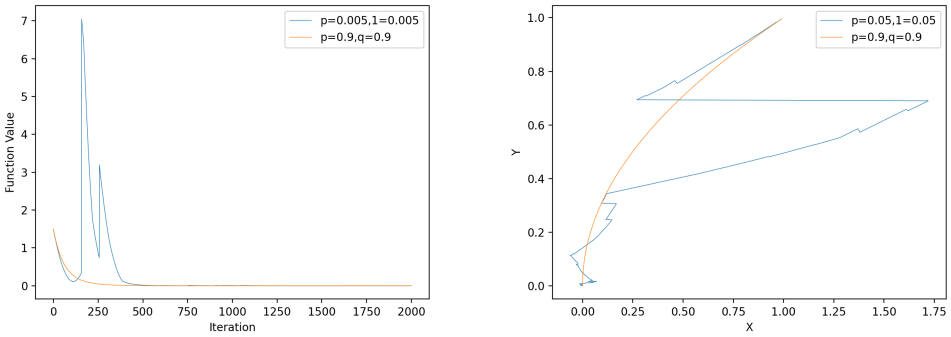
$$V^k \overset{def}{=} \left\|x^k - x^\star\right\|^2 + M\gamma^2\sigma^k$$

Proof. We only need to prove the lemma, the theorem is a corollary of the lemma.

$$\begin{aligned}
&\mathrm{E}\left[||g^k - \nabla f(x^\star)||^2 \mid k\right] = (1-q)||\nabla f(x^k) - \nabla f(x^\star)||^2 \\
&+ q||1/q(\nabla f(x^k) - \nabla f(x^\star)) + (1-1/q)(\nabla f(y^k) - \nabla f(x^\star))||^2 \\
&\leq (1 - q + \frac{2(1-q)^2}{q})||\nabla f(y^k) - \nabla f(x^\star)||^2 + 2(\frac{2L}{q})D_f(x^k, x^\star) \\
\mathrm{E}\left[||\nabla f(y^{k+1}) - \nabla f(x^\star)||^2 \mid k\right] &= (1-p)||\nabla f(y^k) - \nabla f(x^\star)|| + p||\nabla f(x^k) - \nabla f(x^\star)||^2 \\
&\leq (1-p)\sigma^k + 2(pL)D_f(x^k, x^\star).
\end{aligned}$$

□

(iii)when $p \approx 1, q \approx 1$, this method perform well, when $p \approx 0, q \approx 0$, this method performs poorly. This method will reduce the computation of the gradient, so when it is very expensive to compute the gradient(like the dimension is very large), then this method might be a good substitute of gradient descent method. In my experiment, $f = x^2 + 0.5y^2$,step size $\gamma = 0.005$, I get the results in terms of different p, q.



(a) (b)

Figure 1: (a) shows function value change in terms of iteration,(b) shows the trajectories of this method with different p, q.you can see that when p, q are close to 1, this method converges quite smoothly, when p, q are close to 0, this method converges with fluctuations, this match the theory.