CS331-Exam-Lukang-Sun

December 15, 2021

p1. since

$$\frac{d^2}{dt^2} - \log(t) = \frac{1}{t^2} > 0,$$

so $-\log(t)$ is convex on $(0, +\infty)$.

$$\prod_{i=1}^{k} x_i^{\alpha_i} = e^{\sum_{i=1}^{k} \alpha_i \log(x_i)} \leq e^{\log(\sum_{i=1}^{k} \alpha_i x_i)} = \sum_{i=1}^{k} \alpha_i x_i,$$

the first inequality uses the concavity of $\log(t)$ and the increasing property of e^t.

p2. since the log-convexity property of f, we know

$$\log f(\alpha x_1 + \beta x_2) \leq \alpha \log f(x_1) + \beta \log f(x_2),$$

where $\alpha \geq 0, \beta \geq 0, \alpha + \beta = 1$. This is equivalent to

$$f(\alpha x_1 + \beta x_2) \leq f^{\alpha}(x_1) f^{\beta}(x_2). \tag{1}$$

Using Young inequality, the RHS of (1) is bounded by $\alpha f(x_1) + \beta f(x_2)(p = \frac{1}{\alpha}, q = \frac{1}{\beta}$ in this case), so finally proved

$$f(\alpha x_1 + \beta x_2) \leq \alpha f(x_1) + \beta f(x_2).$$

p3. convex since

$$\frac{e^{2021x} - 1}{e^x - 1} = \sum_{i=0}^{2020} e^{ix}$$

$$\left(\frac{e^{2021x} - 1}{e^x - 1}\right)' = \sum_{i=0}^{2020} ie^{ix}$$

$$\left(\frac{e^{2021x} - 1}{e^x - 1}\right)'' = \sum_{i=0}^{2020} i^2 e^{ix}$$

$$(\log(g(t)))'' = \frac{g''g - g'^2}{g^2}$$

so we only need to verify

$$\left(\frac{e^{2021x} - 1}{e^x - 1}\right)'' \frac{e^{2021x} - 1}{e^x - 1} \geq \left(\frac{e^{2021x} - 1}{e^x - 1}\right)'^2$$

. if we expand both side and compare the coefficient in front of $e^{(i+j)x} + e^{(j+i)x}$, we find that LHS is $i^2 + j^2$, RHS is $2ij$ which is smaller than $i^2 + j^2$ by Cauchy-Schwartz inequality, so we proved the LHS is bigger than RHS, that means f is convex.

p4. $(i) \mapsto (ii)$ since $g := \frac{1}{2}||x||_L - f(x)$ is convex, so we have

$$g(y) + \langle \nabla g(y), x - y \rangle \leq g(x),$$

that is

$$1/2||y||_L - f(y) + \langle Ly - \nabla f(y), x - y \rangle \leq 1/2||x||_L - f(x),$$

which is

$$f(x) \leq f(y) + \langle \nabla f(y), x-y \rangle + 1/2||x||_L - 1/2||y||_L - \langle Ly, x-y \rangle = f(y) + \langle \nabla f(y), x-y \rangle + 1/2||x-y||_L.$$

$(ii) \mapsto (iii)$ by (ii), we also have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + 1/2||x-y||_L$$

by summing the last one and (ii), we get (iii).

$(iii) \mapsto (ii)$let $z = y + t(x-y)$

$$\begin{aligned}
f(x) &= f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt \\
&= f(y) + \langle \nabla f(y), x-y \rangle + \int_0^1 \langle \nabla f(z) - \nabla f(y), x - y \rangle dt \\
&\overset{(iii)}{\leq} f(y) + \langle \nabla f(y), x-y \rangle + \int_0^1 t||x-y||_L dt \\
&= f(y) + \langle \nabla f(y), x-y \rangle + 1/2||x-y||_L.
\end{aligned}$$

$(ii) \mapsto (i)$ let $z = \lambda x + (1-\lambda)y$, then by (ii) we have

$$f(x) \leq f(z) + (1-\lambda)\langle \nabla f(z), x-y \rangle + (1-\lambda)^2/2||x-y||_L^2 \tag{2}$$

$$f(y) \leq f(z) + \lambda \langle \nabla f(z), y-x \rangle + \lambda^2/2||x-y||_L^2 \tag{3}$$

(2) times λ plus (3) times $1-\lambda$, we get

$$\lambda f(x) + (1-\lambda) f(y) \leq f(z) + \lambda(1-\lambda)/2||x-y||_L^2 \tag{4}$$

since

$$||z||_L^2 - \lambda||x||_L^2 - (1-\lambda)||y||_L^2 = -\lambda(1-\lambda)\left(||x||_L^2 + ||y||_L^2 - 2\langle x, y \rangle\right) = -\lambda(1-\lambda)||x-y||_L^2,$$

(4) is equivalent to $g(z) \leq \lambda g(x) + (1-\lambda) g(y)$.

p5. let $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$,$x$ is fixed, then we know ϕ is convex so x is the minimum point (convex+linear function is also convex) and we can easily verify that ϕ satisfies (iii) of problem 4, so (ii) is also satisfied, then

$$\phi(x) \leq \phi(y - L^{-1}\nabla\phi(y)) \leq \phi(y) - 1/2||\nabla \phi(y)||_{L^{-1}}^2,$$

that is

$$f(x) + \langle \nabla f(x), y - x \rangle + 1/2||\nabla f(x) - \nabla f(y)||_{L^{-1}}^2 \leq f(y),$$

by changing the position of x, y in the above inequality and take summation of the two inequalities, we finally get

$$||\nabla f(x) - \nabla f(y)||_{L^{-1}}^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

p6. let $s = ty + (1-t)z$

$$\begin{aligned}
\langle s, x \rangle - f(x) &\leq t(\langle y, x \rangle - f(x)) + (1-t)(\langle z, x \rangle - f(x)) \\
&\leq \sup_x \left(t(\langle y, x \rangle - f(x)) + (1-t)(\langle z, x \rangle - f(x))\right) \\
&\leq \sup_x t(\langle y, x \rangle - f(x)) + \sup_x (1-t)(\langle z, x \rangle - f(x)) = tf^*(y) + (1-t)f^*(z)
\end{aligned}$$

then LHS takes supremum, we have $f^*(s) \leq tf^*(y) + (1-t)f^*(z)$, which proves the comvexity.

p7. by assumption, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \le ||\nabla f(x) - \nabla f(y)||||x - y|| \le L||x - y||^2$$

then we know f is L-smoothness(see problem 4), so we have

$$2D_f(x, y) \le L||x - y||_2^2$$

Also we have

$$-\langle \nabla f(x) - \nabla f(y), x - y \rangle \le ||\nabla f(x) - \nabla f(y)||||x - y|| \le L||x - y||^2$$

, so $-f(x)$ is L-smoothness, so we have

$$-2D_f(x, y) = 2D_{-f}(x, y) \le L||x - y||_L^2$$

. So we always have

$$2|D_f(x, y)| \le 2||x - y||^2.$$

p8. for example let $d = n = \lceil \frac{\max_i L_i}{L} \rceil$ and $f_i(x) = \frac{L_i}{2} x_i^2$, then each f_i is convex and L_i-smooth. f is $\frac{\max_i L_i}{n}-$ smooth, $\frac{\max_i L_i}{n} = \frac{\max_i L_i}{\lceil \frac{\max_i L_i}{L} \rceil} \le L$, so f is also $L-$smooth(since L is bigger than the real smoothness constant.)

p9.

$$E\left[\left\|x^k - x^\star\right\|^2\right] \le (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\sigma_\star^2}{\mu}, \tag{5}$$

we want the first term and the second term both less than $\epsilon/2$, so $\gamma = \frac{\epsilon\mu}{4\sigma_*^2}$, $k = \frac{1}{\gamma\mu} \log(\frac{2||x_0 - x_*||^2}{\epsilon})$, the expected computation complexity is

$$k \sum_{i=1}^n \frac{c_i p_i + 1}{n} = \frac{4\sigma_*^2}{\epsilon\mu^2} \log(\frac{2||x_0 - x_*||^2}{\epsilon}) \sum_{i=1}^n \frac{c_i p_i + 1}{n},$$

whichis equivalent to minimize

$$(\sum_{i=1}^n c_i p_i + 1)(\sum_{i=1}^n \frac{||\nabla f_i(x_*)||^2}{p_i}),$$

since we don't know the optimal point x_*, so we should make assumption on $||\nabla f(x_*)||^2$, the proper assumption is make them all equal, so under this assumption, we want to minimize $\sum_{i=1}^n c_i p_i + 1$, $p_i = 1$ for $i = argmax_j c_j, p_i = 0$otherwise(assume each c_i not equal), but this is not an unbiased estimator, thereasonable one is let p_i propotional to $(\frac{1}{c_i})^k$, for some integer n, for example $k = 10$, then $p_i = \frac{c_i^{-10}}{\sum_{i=1}^n c_i^{-10}}$.

p10. (i)

Theorem 1. *assume f is L-smoothness and $\mu-$convex, then*

$$\mathrm{E}[||x^k - x^*||^2] \le (1 - \mu\gamma)||x^0 - x^*||^2$$

where $\gamma \le \frac{1}{(1/\tau(\max_s 1/q_s - 1) + 1)L}$

Proof. let $r^k = x^k - x^*, g^k = C(\nabla f(x^k))$, then by lemma 1, we have

$$\begin{aligned}
\mathrm{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] &\le (1 - \gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f\left(x^k, x^\star\right) + \gamma^2 \mathrm{E}\left[\left\|g^k - \nabla f\left(x^\star\right)\right\|^2 \mid x^k\right] \\
&\le (1 - \gamma\mu)\left\|r^k\right\|^2 - 2\gamma D_f\left(x^k, x^\star\right) + 2\gamma^2 A D_f\left(x^k, x^\star\right) + \gamma^2 C \\
&= (1 - \gamma\mu)\left\|r^k\right\|^2 - 2\gamma(1 - \gamma A) D_f\left(x^k, x^\star\right) + \gamma^2 C \\
&\le (1 - \gamma\mu)\left\|r^k\right\|^2 + \gamma^2 C,
\end{aligned} \tag{6}$$

unrolling the recurence, we get

$$\begin{aligned}
\mathrm{E}\left[\left\|r^k\right\|^2\right] &\leq (1-\gamma\mu)^k \left\|r^0\right\|^2 + \gamma^2 C \sum_{i=0}^{k-1}(1-\gamma\mu)^i \\
&\leq (1-\gamma\mu)^k \left\|r^0\right\|^2 + \frac{\gamma C}{\mu}
\end{aligned} \tag{7}$$

□

Lemma 1. *Assume f is L-smooth,*

$$G^k \overset{def}{=} E\left[\|C(\nabla f(x)) - \nabla f(x^\star)\|^2 \mid x\right] \leq 2AD_f\left(x^k + C, x^\star\right)$$

where $A = (1/\tau(\max_s 1/q_s - 1) + 1)L, C = 0$

Proof.

$$\mathrm{E}[C(x)]\frac{1}{\tau}\sum_{i=1}^{\tau}\mathrm{E}[C_{st}(x)] = x$$

$$\begin{aligned}
\mathrm{E}[||C(x)||^2] &= 1/\tau^2 \sum_{t=1} \mathrm{E}[||C_{st}||^2] + 1/\tau^2 \sum_{i\neq j} \mathbb{E}[C_{si}C_{sj}] \\
&= 1/\tau^2 \sum_{t=1}^{\tau}\sum_{d} 1/q_s x_s^2 + (\tau-1)/\tau||x||^2 \\
&\leq (1/\tau(\max_s 1/q_s - 1) + 1)||x||^2
\end{aligned}$$

so define $w = 1/\tau(\max_s 1/q_s - 1)$, then

$$\begin{aligned}
G(x, x_*) &\overset{\text{def}}{=} \mathrm{E}\left[\|C(\nabla f(x)) - \nabla f(x_*)\|^2\right] \\
&= E\left[\|C(\nabla f(x))\|^2\right] \\
&\leq (\omega+1)\|\nabla f(x) - \nabla f(x_*)\|^2 \\
&\leq 2(\omega+1)LD_f(x, x_*)
\end{aligned}$$

□

(ii) what do you mean the batch size of RCD, do you mean

$$g^k = 1/\tau \sum_{\tau \text{times}} \mathcal{C}_{S^k}\left(\nabla f\left(x^k\right)\right)? \tag{8}$$

if this is the case, then the new one is better, since for each C_{st} it only require computing one coordiante, while general RCD require $|S_k|$ cooridates.The stepsizes are the same.

p11.

Lemma 2. *We have AC-inequality*

$$\mathrm{E}[||g^k - \nabla f(x^*)||^2] \leq 2AD_f(x^k, x^*) + C_k,$$

where $A = (2w+1)L, C_k = 2w||\nabla f(x^*) - h^k||^2 = 2w||h^k||^2.$

Proof.

$$
\begin{aligned}
E\left[\left\|g^k-\nabla f\left(x^{\star}\right)\right\|^2\right] =&E\left[\|g^k-\nabla f(x)\|^2\right]+\|\nabla f(x)-\nabla f\left(x^{\star}\right)\|^2 \\
& \leq \omega||\nabla f(x)-h^k\|^2+\|\nabla f(x)-\nabla f\left(x^{\star}\right)\|^2 \\
=&\omega\left\|\nabla f\left(x^k\right)-\nabla f\left(x^{\star}\right)+\nabla f\left(x^{\star}\right)-h^k\right\|^2+\|\nabla f(x)-\nabla f\left(x^{\star}\right)\|^2 \\
& \stackrel{(35)}{\leq} 2\omega\left\|\nabla f(x)-\nabla f\left(x^{\star}\right)\right\|^2+2\omega\left\|\nabla f\left(x^{\star}\right)-h^k\right\|^2 \\
& +\|\nabla f(x)-\nabla f\left(x^{\star}\right)\|^2 \\
& (2\omega+1)\left\|\nabla f(x)-\nabla f\left(x^{\star}\right)\right\|^2+2\omega\left\|\nabla f\left(x^{\star}\right)-h^k\right\|^2 \\
\leq& 2(2\omega+1)LD_f\left(x,x^{\star}\right)+2\omega\left\|\nabla f\left(x^{\star}\right)-h^k\right\|^2
\end{aligned} \tag{9}
$$

□

Theorem 2. *Assume f is L-smooth,$\mu-$convex, then we have*

$$\mathbb{E}[||x^k-x^*||^k] \leq (1-\gamma\mu)^k||x^0-x^*||^2+\tilde{C}_k$$

where $\gamma \leq \frac{1}{(2w+1)L}, \tilde{C}_k = 2w\sum_{i=0}^{k-1}(1-\gamma\mu)^i||h_{k-1-i}||^2$

Proof. define $r^k = x^k - x^*$, then by lemma 2, we have

$$
\begin{aligned}
\mathrm{E}\left[\left\|r^{k+1}\right\|^2 \mid x^k\right] & \leq(1-\gamma\mu)\left\|r^k\right\|^2-2\gamma D_f\left(x^k,x^{\star}\right)+\gamma^2\mathrm{E}\left[\left\|g^k-\nabla f\left(x^{\star}\right)\right\|^2 \mid x^k\right] \\
& \leq(1-\gamma\mu)\left\|r^k\right\|^2-2\gamma D_f\left(x^k,x^{\star}\right)+2\gamma^2AD_f\left(x^k,x^{\star}\right)+\gamma^2C_k \\
& =(1-\gamma\mu)\left\|r^k\right\|^2-2\gamma(1-\gamma A)D_f\left(x^k,x^{\star}\right)+\gamma^2C_k \\
& \leq(1-\gamma\mu)\left\|r^k\right\|^2+\gamma^2C_k,
\end{aligned} \tag{10}
$$

unrolling the recurence, we get

$$
\begin{aligned}
\mathrm{E}\left[\left\|r^k\right\|^2\right] & \leq(1-\gamma\mu)^k\left\|r^0\right\|^2+\gamma^2\sum_{i=0}^{k-1}C_{k-1-i}(1-\gamma\mu)^i \\
& \leq(1-\gamma\mu)^k\left\|r^0\right\|^2+\frac{\gamma\tilde{C}_k}{\mu}
\end{aligned} \tag{11}
$$

where $\tilde{C}_k = 2w\sum_{i=0}^{k-1}(1-\gamma\mu)^i||h_{k-1-i}||^2$

□

p12.

Lemma 3. *Assume f_i, f are L_i, L smooth. Let*

$$a_i(x,y) \stackrel{def}{=} \nabla f_i(x) - \nabla f_i(y)$$

and

$$\bar{a}(x,y) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}a_i(x,y) = \nabla f(x) - \nabla f(y)$$

Then

$$\mathbb{E}[||\frac{1}{np_i}a_i(x,y)-\bar(a)(x,y)||^2] \leq (max_i\frac{L_i^2}{np_i}+L^2)||x-y||^2$$

Lemma 4. *Suppose that lemma 3 holds. Let*

$$\sigma(x,y) \stackrel{def}{=} \|x-y\|^2.$$

The L-SVRG-NS gradient estimator is unbiased, and for each $\beta > 0$ satisfies the recursions

$$\mathrm{E}\left[\left\|g^k\right\|^2\right] \leq \underbrace{\alpha}_{B_1} \mathrm{E}\left[\sigma^k\right] + \underbrace{1}_{B_2} \mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$$

$$\mathrm{E}\left[\sigma^{k+1}\right] \leq \underbrace{(1-p)\left(1+\gamma\beta+\gamma^2\alpha\right)}_{\tilde{B}_1} \mathrm{E}\left[\sigma^k\right] + \underbrace{(1-p)\left(\gamma\beta^{-1}+\gamma^2\right)}_{\tilde{B}_2} \mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$$

where $\alpha \overset{def}{=} \max_i \frac{L_i^2}{p_i} + L^2$ *and* $\sigma^k \overset{def}{=} \sigma\left(x^k, y^k\right) = \left\|x^k - y^k\right\|^2$.

Proof.

$$a_i(x,y) \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y)$$

and $\bar{a} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n a_i(x,y) = \nabla f(x) - \nabla f(y)$, then by variance decomposition we can write

$$\begin{aligned}
\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] &= \mathrm{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2 \mid x^k, y^k\right] + \left\|\nabla f\left(x^k\right)\right\|^2 \\
&= \mathrm{E}\left[\left\|\frac{\nabla f_i(x^k)}{np_i} - \frac{\nabla f_i(y^k)}{np_i} + \nabla f\left(y^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k, y^k\right] + \left\|\nabla f\left(x^k\right)\right\|^2 \\
&= \mathbb{E}\left[\left\|\frac{1}{np_i} a_i(x^k, y^k) - \bar{a}\left(x^k, y^k\right)\right]\right\|^2 \mid x^k, y^k + \left\|\nabla f\left(x^k\right)\right\|^2 \\
&\leq (max_i \frac{L_i^2}{np_i} + L^2)\sigma^k + \left\|\nabla f\left(x^k\right)\right\|^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] &= \left[\left\|x^{k+1} - y^{k+1}\right\|^2 \mid x^{k+1}, x^k, y^k\right] \\
&= (p)\left\|x^{k+1} - x^{k+1}\right\|^2 + (1-p)\left\|x^{k+1} - y^k\right\|^2 \\
&= (1-p)\left\|x^{k+1} - x^k + x^k - y^k\right\|^2 \\
&= (1-p)\left\|x^{k+1} - x^k\right\|^2 + 2(1-p)\left\langle x^{k+1} - x^k, x^k - y^k\right\rangle + (1-p)\left\|x^k - y^k\right\|^2 \\
&= (1-p)\gamma^2\left\|g^k\right\|^2 + 2(1-p)\left\langle x^{k+1} - x^k, x^k - y^k\right\rangle + (1-p)\sigma^k \\
&= (1-p)\gamma^2\left\|g^k\right\|^2 + 2(1-p)\gamma\left\langle -g^k, x^k - y^k\right\rangle + (1-p)\sigma^k
\end{aligned} \tag{12}$$

Now taking conditional expectation again, but this time conditioning on x^k and y^k only, and applying the inequality

$$\langle u, v\rangle \leq \frac{1}{2\beta}\|u\|^2 + \frac{\beta}{2}\|v\|^2$$

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] \mid x^k, y^k\right] \leq &(1-p)\gamma^2\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + 2(1-p)\gamma\langle\underbrace{\mathrm{E}\left[g^k \mid x^k, y^k\right]}_{\nabla f(x^k)}, y^k - x^k\rangle \\
&+(1-p)\sigma^k \\
&\leq (1-p)\gamma^2\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + 2(1-p)\gamma\left(\frac{1}{2\beta}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\beta}{2}\left\|x^k - y^k\right\|^2\right) \\
&\quad + (1-p)\sigma^k \\
&= (1-p)\gamma^2\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + \frac{(1-p)\gamma}{\beta}\left\|\nabla f\left(x^k\right)\right\|^2 + (1-p)(1+\gamma\beta)\sigma^k.
\end{aligned} \tag{13}$$

For simplicity, in what follows we denote $\alpha \overset{\text{def}}{=} \max_i \frac{L_i^2}{p_i} + L^2$. Applying expectation one more time, and using the more elaborate tower property of expectation (335), we get

$$
\begin{aligned}
\mathrm{E}\left[\sigma^{k+1}\right] =& \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[\sigma^{k+1} \mid x^{k+1}, x^k, y^k\right] \mid x^k, y^k\right]\right] \\
=& \mathrm{E}\left[(1-p)\gamma^2 \mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right] + \frac{(1-p)\gamma}{\beta}\left\|\nabla f\left(x^k\right)\right\|^2 + (1-p)(1+\gamma\beta)\sigma^k\right] \\
=& (1-p)\gamma^2 \mathrm{E}\left[\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, y^k\right]\right] + \frac{(1-p)\gamma}{\beta}\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathrm{E}\left[\sigma^k\right] \\
=& (1-p)\gamma^2 \mathrm{E}\left[\alpha\sigma^k + \left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{(1-p)\gamma}{\beta}\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathrm{E}\left[\sigma^k\right] \\
=& (1-p)\gamma^2\alpha \mathrm{E}\left[\sigma^k\right] + (1-p)\gamma^2 \mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{(1-p)\gamma}{\beta}\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + (1-p)(1+\gamma\beta)\mathrm{E}\left[\sigma^k\right] \\
=& (1-p)\left(1+\gamma\beta+\gamma^2\alpha\right)\mathrm{E}\left[\sigma^k\right] + (1-p)\left(\gamma\beta^{-1}+\gamma^2\right)\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right].
\end{aligned}
$$

□

Remark 1.

$$
\frac{1}{\gamma} \geq \sqrt{\frac{4}{3}\frac{1-p}{p}\alpha(c+1)}
$$

then

$$
\tilde{B}_1 \leq 1 - \frac{p}{4}
$$

Indeed,

$$
\begin{aligned}
\tilde{B}_1 =& (1-p)\left(1+\gamma\beta+\gamma^2\alpha\right) \\
=& (1-p)\left(1+\gamma^2\alpha(c+1)\right) \\
=& 1-p+\gamma^2\alpha(c+1)(1-p) \\
\leq& 1-p+\frac{3p}{4} \\
=& 1-\frac{p}{4}.
\end{aligned}
$$

Theorem 3. *Let f_i, f are $L_i, L-$ smooth and . Choose constant stepsize γ satisfying*

$$
0 < \gamma \leq \frac{1}{L\left(B_2 + \theta\tilde{B}_2\right)}
$$

where $\theta \overset{def}{=} \frac{B_1}{1-\tilde{B}_1}$. Then for any $K \geq 1$, SGD-CTRL can output a random point x (chosen as one of the points $x^0, x^1, \ldots, x^{K-1}$ at random with certain probabilities) satisfying

$$
\mathrm{E}\left[\|\nabla f(x)\|^2\right] \leq L(C+\theta\tilde{C})\gamma + \frac{2\left(1+L(A+\theta\tilde{A})\gamma^2\right)^K}{\gamma K}\Delta^0
$$

where $\Delta^0 \overset{def}{=} f\left(x^0\right) - f^{\inf} + \frac{1}{2}L\theta\sigma^0\gamma^2$, all the constants are from lemma 4.

Proof. Since f is L-smooth, we have

$$
\begin{aligned}
f\left(x^{k+1}\right) - f^{\inf} \leq& f\left(x^k\right) - f^{\inf} + \left\langle\nabla f\left(x^k\right), x^{k+1}-x^k\right\rangle + \frac{L}{2}\left\|x^{k+1}-x^k\right\|^2 \\
=& f\left(x^k\right) - f^{\inf} - \gamma\left\langle\nabla f\left(x^k\right), g^k\right\rangle + \frac{L\gamma^2}{2}\left\|g^k\right\|^2
\end{aligned}
$$

By applying expectation to both sides and subsequently using unbiasedness of g^k and the bound lemma 4 on the second moment of the stochastic gradient, we get

$$
\begin{aligned}
&E\left[f\left(x^{k+1}\right)-f^{\inf} \mid x^k, \xi^k\right] \\
&\leq \; f\left(x^k\right)-f^{\inf}-\gamma\left\|\nabla f\left(x^k\right)\right\|^2+\tfrac{L\gamma^2}{2}\mathrm{E}\left[\left\|g^k\right\|^2 \mid x^k, \xi^k\right] \\
&\leq \; f\left(x^k\right)-f^{\inf}-\gamma\left\|\nabla f\left(x^k\right)\right\|^2 \\
&+\tfrac{L\gamma^2}{2}\left[2A\left(f\left(x^k\right)-f^{\inf}\right)+B_1\sigma^k+B_2\left\|\nabla f\left(x^k\right)\right\|^2+C\right] \\
&\qquad =\left(1+LA\gamma^2\right)\left(f\left(x^k\right)-f^{\inf}\right)+\frac{LB_1\gamma^2}{2}\sigma^k \\
&\qquad -\left(\gamma-\frac{LB_2\gamma^2}{2}\right)\left\|\nabla f\left(x^k\right)\right\|^2+\frac{LC\gamma^2}{2}
\end{aligned}
$$

Choose any $M>0$ and define

$$
\Delta^{k+1} \stackrel{\text{def}}{=} f\left(x^{k+1}\right)-f^{\inf}+M\gamma^2\sigma^{k+1}
$$

we get

$$
\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right] \leq \underbrace{\left(1+LA\gamma^2+2M\tilde{A}\gamma^2\right)}_{=a}\left(f\left(x^k\right)-f^{\inf}\right)+\left(\frac{LB_1}{2}+M\tilde{B}_1\right)\gamma^2\sigma^k
$$

$$
-\underbrace{\left(\gamma-\frac{LB_2\gamma^2}{2}-M\tilde{B}_2\gamma^2\right)}_{=b}\left\|\nabla f\left(x^k\right)\right\|^2+\underbrace{\frac{LC\gamma^2}{2}+M\tilde{C}\gamma^2}_{=c} a\left[f\left(x^k\right)-f^{\inf}+\frac{\frac{LB_1}{2}+M\tilde{B}_1}{a}\gamma^2\sigma^k\right]-b\left\|\nabla f\left(x^k\right)\right\|^2+c
$$

,

where

$$
\begin{aligned}
a &\stackrel{\text{def}}{=} 1+LA\gamma^2+2M\tilde{A}\gamma^2 \\
b &\stackrel{\text{def}}{=} \gamma-\frac{LB_2\gamma^2}{2}-M\tilde{B}_2\gamma^2 \\
c &\stackrel{\text{def}}{=} \frac{LC\gamma^2}{2}+M\tilde{C}\gamma^2.
\end{aligned}
$$

In order to turn the last inequality into a recursion which has Δ^k on the right hand side, we need to make sure that

$$
\frac{\frac{LB_1}{2}+M\tilde{B}_1}{a} \leq M.
$$

Fortunately, it is easy to see (prove this!) that we can make sure this holds by an appropriate choice of M. In particular,the last inequality holds if we choose

$$
M \stackrel{\text{def}}{=} \frac{LB_1}{2\left(1-\tilde{B}_1\right)}=\frac{L\theta}{2}
$$

With this choice of M, we can obtain the recursion

$$
\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right] \leq a\Delta^k-b\left\|\nabla f\left(x^k\right)\right\|^2+c.
$$

By applying expectation to both sides of this, and using the tower property of expectation, we get the recursion

$$
\begin{aligned}
\mathrm{E}\left[\Delta^{k+1}\right] &= \mathrm{E}\left[\mathrm{E}\left[\Delta^{k+1} \mid x^k, \xi^k\right]\right] \\
&\leq a\mathrm{E}\left[\Delta^k\right]-b\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]+c.
\end{aligned}
$$

We now apply Lemma 120 from lecturn to recursion with $X_k=E\left[\Delta^k\right]$ and $Y_k=bE\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$. If we set $x=x^k$ with probability p_k (where p_k is as in Lemma 120), which means that $Y=Y_k$ with probability p_k, we conclude that

$$
\begin{aligned}
bE\left[\|\nabla f(x)\|^2\right] &= E[Y] \\
&\leq \frac{a^K}{S_K}\Delta^0 + c \\
&\leq \frac{a^K}{K}\Delta^0 + c
\end{aligned}
$$

where the last inequality follows since $a \geq 1$, which implies that $S_K \geq K$. We now evaluate the expressions for b and c in (317). First,

$$
\begin{aligned}
b=&\gamma - \frac{LB_2\gamma^2}{2} - M\tilde{B}_2\gamma^2 \\
=&\gamma - \frac{\gamma}{2}\left(LB_2\gamma + L\theta\tilde{B}_2\gamma\right) \\
\geq& \frac{\gamma}{2}
\end{aligned}
$$

where the last inequality holds by setting

$$
\gamma \leq \frac{1}{L\left(B_2 + \theta\tilde{B}_2\right)}
$$

Moreover,

$$
c= \quad \frac{LC}{2}\gamma^2 + M\tilde{C}\gamma^2= \quad \frac{L}{2}(C + \theta\tilde{C})\gamma^2.
$$

We obtain the results. □

p13.

Lemma 5. $g(x) = 1/n\sum_{i=1}^{n}\mathcal{C}_i(\nabla f_i(x))$, *assume* $\mathbb{E}[||g(x) - \nabla f(x)||^2] \leq C$, f *is L-smooth, then*

$$
E\left[\|g(x)\|^2\right] \leq 2A\left(f(x) - f^{\inf}\right) + B\|\nabla f(x)\|^2 + C, \tag{14}
$$

where $A = L, B = 0$

Proof.

$$
\begin{aligned}
\mathbb{E}[||g(x)||^2] &= \mathbb{E}[||g(x) - \nabla f(x)||^2] + ||\nabla f(x) - \nabla f(x^*)(=0)||^2 \\
&\leq 2L(f(x) - f^{inf}) + C
\end{aligned}
$$

□

Lemma 6. *Assume the conditions in lemma 5 hold true. Choose constant stepsize* γ *satisfying* $0 < \gamma \leq \frac{1}{LB}$. *Then for any* $K \geq 1$, *the iterates* $\left\{x^k\right\}$ *of SGD satisfy*

$$
\frac{1}{2}\sum_{k=0}^{K-1} w_k r^k + \frac{w_{K-1}}{\gamma}\delta^K \leq \frac{w_{-1}}{\gamma}\delta^0 + \frac{LC}{2}\sum_{k=0}^{K-1} w_k\gamma.
$$

where $r^k \stackrel{def}{=} \mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right], w_k \stackrel{def}{=} \frac{w_{-1}}{(1+LA\gamma^2)^{k+1}}$ *for* $w_{-1} > 0$ *arbitrary, and* $\delta^k \stackrel{def}{=} \mathrm{E}\left[f\left(x^k\right)\right] - f^{inf}$.

Proof. We start with the L-smoothness of f, which implies

$$
\begin{aligned}
f\left(x^{k+1}\right) \leq& f\left(x^k\right) + \left\langle\nabla f\left(x^k\right), x^{k+1} - x^k\right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\
=&f\left(x^k\right) - \gamma\left\langle\nabla f\left(x^k\right), g\left(x^k\right)\right\rangle + \frac{L\gamma^2}{2}\left\|g\left(x^k\right)\right\|^2
\end{aligned}
$$

Step 2: Applying the (ABC) Assumption. Taking expectation conditional on x^k, and using lemma 5, we get

$$\begin{aligned}
\mathrm{E}\left[f\left(x^{k+1}\right) \mid x^k\right] &= f\left(x^k\right)-\gamma\left\|\nabla f\left(x^k\right)\right\|^2+\frac{L\gamma^2}{2}\mathrm{E}\left[\left\|g\left(x^k\right)\right\|^2\right] \\
&\overset{\text{(ABC)}}{\leq} \quad f\left(x^k\right)-\gamma\left\|\nabla f\left(x^k\right)\right\|^2+\frac{L\gamma^2}{2}\left(2A\left(f\left(x^k\right)-f^{\inf}\right)+B\left\|\nabla f\left(x^k\right)\right\|^2+C\right) \\
&= f\left(x^k\right)-\gamma\left(1-\frac{LB\gamma}{2}\right)\left\|\nabla f\left(x^k\right)\right\|^2+LA\gamma^2\left(f\left(x^k\right)-f^{\inf}\right)+\frac{LC\gamma^2}{2}.
\end{aligned}$$

Subtracting $f^{\inf}$ from both sides gives $\mathrm{E}\left[f\left(x^{k+1}\right) \mid x^k\right]-f^{\inf} \leq \left(1+LA\gamma^2\right)\left(f\left(x_k\right)-f^{\inf}\right)-\gamma\left(1-\frac{LB\gamma}{2}\right)\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{LC\gamma^2}{2}$ taking expectation again, using the tower property and rearranging, we get $\mathrm{E}\left[f\left(x^{k+1}\right)-f^{\inf}\right]+\gamma\left(1-\frac{LB\gamma}{2}\right)\mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leq \left(1+LA\gamma^2\right)\mathrm{E}\left[f\left(x^k\right)-f^{\inf}\right]+\frac{LC\gamma^2}{2}$. Letting $\delta^k \overset{\text{def}}{=} \mathrm{E}\left[f\left(x^k\right)-f^{\inf}\right]$ and $r^k \overset{\text{def}}{=} \mathrm{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$, we can rewrite the last inequality as

$$\gamma\left(1-\frac{LB\gamma}{2}\right)r^k \leq \left(1+LA\gamma^2\right)\delta^k-\delta^{k+1}+\frac{LC\gamma^2}{2}$$

Our choice of stepsize guarantees that $1-\frac{LB\gamma}{2} \geq \frac{1}{2}$. As such,

$$\frac{\gamma}{2}r^k \leq \left(1+LA\gamma^2\right)\delta^k-\delta^{k+1}+\frac{LC\gamma^2}{2} \tag{15}$$

for $k \geq 0$. We now define an exponentially decaying weighting sequence $w_0, w_1, w_2, \ldots, w_K$. We are interested in the weighting sequence solely as a proof technique, and it does not show up in the final bounds. Fix $w_{-1} > 0$ and define

$$w_k = \frac{w_{k-1}}{1+LA\gamma^2} \qquad \text{for all} \qquad k \geq 0$$

Multiplying recursion (15) by $\frac{w_k}{\gamma}$, we get

$$\begin{aligned}
\frac{1}{2}w_k r^k &\leq \frac{w_k\left(1+LA\gamma^2\right)}{\gamma}\delta^k-\frac{w_k}{\gamma}\delta^{k+1}+\frac{LC\gamma w_k}{2} \\
&= \frac{w_{k-1}}{\gamma}\delta^k-\frac{w_k}{\gamma}\delta^{k+1}+\frac{LC\gamma w_k}{2}
\end{aligned}$$

Summing up both sides for $k = 0, 1, \ldots, K-1$, and noticing that many terms telescope, we get

$$\frac{1}{2}\sum_{k=0}^{K-1} w_k r^k \leq \frac{w-1}{\gamma}\delta^0-\frac{w_{K-1}}{\gamma}\delta^K+\frac{LC\gamma}{2}\sum_{k=0}^{K-1} w_k.$$

Rearranging we get the lemma's statement. □

Theorem 4. *Assume the assumption of lemma 5 holds, then*

$$\min_{0 \leq k \leq K-1} E\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leq LC\gamma+\frac{2\left(1+LA\gamma^2\right)^K}{\gamma K}\delta^0$$

where $\delta^0 \overset{\text{def}}{=} f\left(x^0\right)-f^{\inf}$.

Proof. We start with Lemma 6, which says that

$$\frac{1}{2}\sum_{k=0}^{K-1} w_k r^k \leq \frac{1}{2}\sum_{k=0}^{K-1} w_k r^k+\frac{w_{k-1}}{\gamma}\delta^K \overset{(271)}{\leq} \frac{w_{-1}}{\gamma}\delta^0+\frac{LC}{2}\sum_{k=0}^{K-1} w_k\gamma$$

Let $W_K \stackrel{\text{def}}{=} \sum_{k=0}^{K-1} w_k$. Dividing both sides by W_K, we obtain

$$\frac{1}{2}\min_{0\leq k\leq K-1} r^k \leq \frac{1}{2W_k}\sum_{k=0}^{k-1} w_k r^k \leq \frac{w_{-1}}{W_k}\frac{\delta^0}{\gamma} + \frac{LC\gamma}{2}.$$

Note that

$$W_K = \sum_{k=0}^{K-1} w_k \geq \sum_{k=0}^{K-1} \min_{0\leq i\leq K-1} w_i = K w_{K-1} = \frac{K w_{-1}}{\left(1+LA\gamma^2\right)^K}.$$

Using this in (272) yields

$$\frac{1}{2}\min_{0\leq k\leq K-1} r^k \leq \frac{\left(1+LA\gamma^2\right)^K}{\gamma K}\delta^0 + \frac{LC\gamma}{2}.$$

Multiplying both sides by 2 yields the theorem's claim. $\square$