SGD Ex

Beryl Aribowo

July 13, 2022

[P4]

E1

Def. 17:

$$D_f(x,y) + D_f(y,x) = \langle \nabla f(x) - \nabla f(y), x - y \rangle = \langle \nabla f(y) - \nabla f(x), y - x \rangle \tag{1}$$

$\forall x, y \in \mathbb{R}^d$:

$$
\begin{aligned}
\mu \|x-y\|^2 &\leq 2D_f(x,y), \\
\frac{\mu}{2}\|x-y\|^2 &\leq D_f(x,y), \\
\frac{\mu}{2}\|x-y\|^2 &\leq D_f(y,x), \\
D_f(x,y) + \frac{\mu}{2}\|x-y\|^2 &\leq D_f(x,y) + D_f(y,x), \\
D_f(x,y) + \frac{\mu}{2}\|x-y\|^2 &\overset{(1)}{\leq} \langle \nabla f(x) - \nabla f(y), x-y \rangle.
\end{aligned}
\tag{2}
$$

E2

$$
\begin{aligned}
D_f(x,y) + \frac{\mu}{2}\|x-y\|^2 &\leq \langle \nabla f(x) - \nabla f(y), x-y \rangle, \\
\langle \nabla f(x) - \nabla f(y), x-y \rangle &\geq \underbrace{D_f(x,y)}_{\geq \frac{\mu}{2}\|x-y\|^2} + \frac{\mu}{2}\|x-y\|^2, \\
\langle \nabla f(x) - \nabla f(y), x-y \rangle &\geq \frac{\mu}{2}\|x-y\|^2 + \frac{\mu}{2}\|x-y\|^2, \\
\langle \nabla f(x) - \nabla f(y), x-y \rangle &\geq \mu\|x-y\|^2.
\end{aligned}
\tag{3}
$$

[P6]

E17

(Equation 34):

$$\langle a, b \rangle \leq \frac{||a||^2}{2t} + \frac{t||b||^2}{2},$$
$$\langle a, b \rangle \leq \frac{\langle a, a \rangle}{2t} + \frac{t\langle b, b \rangle}{2},$$
$$2t\langle a, b \rangle \leq \langle a, a \rangle + t^2\langle b, b \rangle,$$
$$0 \leq \langle a, a \rangle + \langle tb, tb \rangle - \langle a, tb \rangle - \langle tb, a \rangle,$$
$$0 \leq ||a - tb||^2.$$

$$(4)$$

(Equation 35):

$$||a + b||^2 \leq 2||a||^2 + 2||b||^2,$$
$$\langle a, a \rangle + \langle b, b \rangle + 2\langle a, b \rangle \leq 2\langle a, a \rangle + 2\langle b, b \rangle,$$
$$0 \leq \langle a, a \rangle + \langle b, b \rangle - 2\langle a, b \rangle,$$
$$0 \leq ||a - b||^2.$$

$$(5)$$

(Equation 36):

$$\frac{1}{2}||a||^2 - ||b||^2 \leq ||a + b||^2,$$
$$\frac{1}{2}\langle a, a \rangle - \langle a, a \rangle \leq \langle a, a \rangle + \langle b, b \rangle + 2\langle a, b \rangle,$$
$$\langle a, a \rangle - 2\langle b, b \rangle \leq 2\langle a, a \rangle + 2\langle b, b \rangle + 4\langle a, b \rangle,$$
$$0 \leq \langle a, a \rangle + \langle 2b, 2b \rangle + \langle a, 2b \rangle + \langle 2b, a \rangle,$$
$$0 \leq ||a + 2b||^2.$$

$$(6)$$

E19

For random vector $X \in \mathbb{R}^d$:

$$\mathbf{Var}[X] := \mathbf{E}\left[||X - \mathbf{E}[X]||^2\right]. \tag{7}$$

Markov's inequality:

$$\mathrm{Prob}(X \geq t) \leq \frac{\mathbf{E}[X]}{t}. \tag{8}$$

Proof of Chebyshev's inequality using Markov's inequality:

$$\mathrm{Prob}(||X - \mathbf{E}[X]||^2 \geq t^2) \leq \frac{\mathbf{E}\left[||X - \mathbf{E}[X]||^2\right]}{t^2}.$$

Since

$$\mathrm{Prob}(||X - \mathbf{E}[X]||^2 \geq t^2) = \mathrm{Prob}(||X - \mathbf{E}[X]|| \geq t), \tag{9}$$

then

$$\mathrm{Prob}(||X - \mathbf{E}[X]|| \geq t) \leq \frac{\mathbf{Var}[X]}{t^2}. \tag{10}$$

[P7]

E24

If
$$f = \frac{1}{n} \sum_{i=1}^{n} f_i,$$

then

$$D_f(x, y) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) - \frac{1}{n} \sum_{i=1}^{n} f_i(y) - \frac{1}{n} \sum_{i=1}^{n} \langle \nabla f_i(y), x - y \rangle,$$

$$D_f(x, y) = \frac{1}{n} \sum_{i=1}^{n} (f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle),$$

$$D_f(x, y) = \frac{1}{n} \sum_{i=1}^{n} D_{f_i}(x, y).$$

E26

If $\sigma_\star^2 = 0$, then

$$\sigma_\star^2 = \left(\frac{1}{n^2} \sum_{i=1}^{n} \frac{||\nabla f_i(x^\star)||^2}{p_i} \right) - ||\nabla f(x^\star)||^2 = 0$$

$$= \left(\frac{1}{n^2} \sum_{i=1}^{n} \frac{||n p_i \nabla f(x^\star)||^2}{p_i} \right) - ||\nabla f(x^\star)||^2 = 0$$

$$= p_i \sum_{i=1}^{n} (||\nabla f(x^\star)||^2) - ||\nabla f(x^\star)||^2 = 0$$

$$= n p_i ||\nabla f(x^\star)||^2 - ||\nabla f(x^\star)||^2 = 0,$$

$$\sigma_\star^2 = 0 \implies n p_i \nabla f(x^\star) = \nabla f(x^\star).$$

[P8]

E33

Let
$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

Since
$$p_i = \frac{1}{n},$$

and
$$|S| = \tau,$$

then
$$\mathbf{E}[\chi_i] = \mathrm{Prob}(i \in S) = \sum_{i=1}^{n} p_i \chi_i = \frac{1}{n} \sum_{i=1}^{n} \chi_i = \frac{\tau}{n}.$$

E35

For any vectors, $b_1, ..., b_n \in \mathbb{R}^d$:

$$\left\|\sum_{i=1}^{n} b_i\right\|^2 - \sum_{i=1}^{n} \|b_i\|^2 = \underbrace{\sum_{i=1}^{n} \langle b_i, b_i \rangle + \sum_{i \neq j} \langle b_i, b_j \rangle - \sum_{i=1}^{n} \langle b_i, b_i \rangle}_{\left\|\sum_{i=1}^{n} b_i\right\|^2},$$

$$\left\|\sum_{i=1}^{n} b_i\right\|^2 - \sum_{i=1}^{n} \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle.$$

[P9]

E37

Assumptions of $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$:

1. $\mathbf{E}[\mathcal{C}(x)] = x, \quad \forall x \in \mathbb{R}^d$

2. $\mathbf{E}\left[||\mathcal{C}(x) - x||^2\right] \leq \omega ||x||^2 + \delta, \quad \forall x \in \mathbb{R}^d, \quad \exists \omega, \delta \geq 0$

Proof of convergence for CGD with $n = 1$:
Since $\mathcal{C} \in \mathbb{B}^d(\omega)$,

$$\mathbf{E}\left[||g(x)||^2\right] = \mathbf{E}\left[||\mathcal{C}(\nabla f(x))||^2\right] \leq (\omega + 1)||\nabla f(x)||^2. \tag{11}$$

In case of $\nabla f(y) = 0$,

$$\begin{aligned}
G(x, y) &:= \mathbf{E}\left[||g(x) - \nabla f(y)||^2\right] \\
&= \mathbf{E}\left[||g(x)||^2\right] \\
&\overset{(11)}{\leq} (\omega + 1)||\nabla f(x) - \nabla f(y)||^2, \\
&\leq 2(\omega + 1)LD_f(x, y).
\end{aligned}$$

In case of $\nabla f(y) \neq 0$,

$$\begin{aligned}
G(x, y) &:= \mathbf{E}\left[||g(x) - \nabla f(y)||^2\right] \\
&= \mathbf{E}\left[||g(x) - \nabla f(x)||^2\right] + ||\nabla f(x) - \nabla f(y)||^2 \\
&= \mathbf{E}\left[||\mathcal{C}(\nabla f(x)) - \nabla f(x)||^2\right] + ||\nabla f(x) - \nabla f(y)||^2 \\
&\leq \omega ||\nabla f(x)||^2 + \delta + ||\nabla f(x) - \nabla f(y)||^2 \\
&= \omega ||\nabla f(x) - \nabla f(y) + \nabla f(y)||^2 + ||\nabla f(x) - \nabla f(y)||^2 + \delta \\
&\leq 2\omega ||\nabla f(x) - \nabla f(y)||^2 + 2\omega ||\nabla f(y)||^2 + ||\nabla f(x) - \nabla f(y)||^2 + \delta \\
&= (2\omega + 1)||\nabla f(x) - \nabla f(y)||^2 + 2\omega ||\nabla f(y)||^2 + \delta \\
&\leq 2 \underbrace{(2\omega + 1)L}_{A} D_f(x, y) + \underbrace{2\omega ||\nabla f(y)||^2 + \delta}_{C}.
\end{aligned}$$

If $0 < \gamma < \frac{1}{A}$, then

$$\mathbf{E}\left[||x^k - x^*||^2\right] \leq (1 - \gamma \mu)^k ||x^0 - x^*|| + \frac{2\gamma \omega ||\nabla f(x^*)||^2 + \gamma \delta}{\mu}.$$

4

E39

Lemma 51:

if $\mathcal{C}(x) = x, \forall x$ (no master compression) and $\omega_i = \omega, \forall i$, then

$$G(x, y) \leq 2 \underbrace{\left(L + 2L_{\max} \frac{\omega}{n} \right)}_{A} D_f(x, y) + \underbrace{2 \frac{\omega}{n} \sigma^2(y)}_{C(y)},$$

where

$$\sigma^2(y) := \frac{1}{n} \sum_{i=1}^{n} ||\nabla f_i(y)||^2.$$

If $\sigma^2(y) = 0$, then

$$G(x, y) \leq 2 \underbrace{\left(L + L_{\max} \frac{\omega}{n} \right)}_{A} D_f(x, y).$$

Proof:

If $\nabla f(y) \neq 0$, then

$$
\begin{aligned}
G(x, y) &:= \mathbf{E} \left[||g(x) - \nabla f(y)||^2 \right] \\
&= \mathbf{E} \left[||g(x) - \nabla f(x)||^2 \right] + ||\nabla f(x) - \nabla f(y)||^2 \quad (12) \\
&\leq \mathbf{E} \left[||g(x) - \nabla f(x)||^2 \right] + 2L D_f(x, y),
\end{aligned}
$$

and

$$g(x) = \mathcal{C}(\hat{g}(x)) = \hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} g_i(x). \quad (13)$$

where

$$g_i(x) = \mathcal{C}_i(\nabla f_i(x)).$$

Estimate

$$
\begin{aligned}
\mathbf{E} \left[||g(x) - \nabla f(x)||^2 \right] &\overset{(13)}{=} \mathbf{E} \left[||\mathcal{C}(\hat{g}(x)) - \nabla f(x)||^2 \right] \\
&= \mathbf{E} \left[||\hat{g}(x) - \nabla f(x)||^2 \right] \\
&= \mathbf{E} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \underbrace{(g_i(x) - \nabla f_i(x))}_{a_i} \right\|^2 \right] \\
&= \frac{1}{n^2} \mathbf{E} \left[\sum_{i=1}^{n} ||a_i||^2 + \sum_{i \neq j} \langle a_i, a_j \rangle \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E} \left[||a_i||^2 \right] + \sum_{i \neq j} \mathbf{E} \left[\langle a_i, a_j \rangle \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E} \left[||a_i||^2 \right] + \sum_{i \neq j} \langle \underbrace{\mathbf{E} \left[a_i \right]}_{0}, \underbrace{\mathbf{E} \left[a_j \right]}_{0} \rangle \\
&\leq \frac{1}{n^2} \sum_{i=1}^{n} \omega_i ||\nabla f_i(x)||^2 \\
&= \frac{\omega}{n^2} \sum_{i=1}^{n} ||\nabla f_i(x)||^2.
\end{aligned}
$$

5

Next, bound
$$||\nabla f_i(x)||^2 = ||\nabla f_i(x) - \nabla f_i(y) + \nabla f_i(y)||^2$$
$$\leq 2||\nabla f_i(x) - \nabla f_i(y)||^2 + 2||\nabla f_i(y)||^2$$
$$\leq 4L_i D_{f_i}(x,y) + 2||\nabla f_i(y)||^2.$$

Combine everything:

$$G(x,y) \leq \mathbf{E}\left[||g(x) - \nabla f(x)||^2\right] + 2LD_f(x,y)$$
$$\leq \frac{\omega}{n^2} \sum_{i=1}^{n} ||\nabla f_i(x)||^2 + 2LD_f(x,y)$$
$$\leq \frac{\omega}{n^2} \sum_{i=1}^{n} \left(4L_i D_{f_i}(x,y) + 2||\nabla f_i(y)||^2\right) + 2LD_f(x,y)$$
$$= 2\frac{\omega}{n}\left(2\sum_{i=1}^{n}\frac{1}{n}L_i D_{f_i}(x,y) + \frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(y)||^2\right) + 2LD_f(x,y) \tag{14}$$
$$\leq 2\frac{\omega}{n}\left(2L_{\max}D_f(x,y) + \sigma^2(y)\right) + 2LD_f(x,y)$$
$$= 2(L + 2L_{\max})D_f(x,y) + 2\frac{\omega}{n}\sigma^2(y).$$

Else, if $\nabla f(y) = 0$, then

$$G(x,y) = \mathbf{E}\left[||\hat{g}(x)||^2\right]$$
$$= \mathbf{E}\left[||\hat{g}(x) - \mathbf{E}\left[\hat{g}(x)\right]||^2\right] + ||\mathbf{E}\left[\hat{g}(x)\right]||^2$$
$$= \mathbf{E}\left[||\hat{g}(x) - \nabla f(x)||^2\right] + ||\nabla f(x)||^2$$
$$\leq \left(\frac{\omega}{n^2}\sum_{i=1}^{n}||\nabla f_i(x)||^2\right) + ||\nabla f(x) - \nabla f(y) + \nabla f(y)||^2$$
$$\leq \left(\frac{\omega}{n^2}\sum_{i=1}^{n}||\nabla f_i(x)||^2\right) + 2||\nabla f(x) - \nabla f(y)||^2 + 2||\underbrace{\nabla f(y)}_{0}||^2 \tag{15}$$
$$\leq \left(\frac{\omega}{n^2}\sum_{i=1}^{n}||\nabla f_i(x)||^2\right) + 2LD_f(x,y)$$
$$\leq \dots \text{ same as (14), from the third line}$$
$$= 2(L + 2L_{\max})D_f(x,y) + 2\frac{\omega}{n}\sigma^2(y).$$

[P10]

E41

Let
$$p_i = \text{Prob}(i \in S),$$
where
$$S \subseteq \{1, 2, ..., d\},$$
then
$$\mathbf{E}[|S|] = \mathbf{E}\left[\sum_{i=1}^{d}|S_i|\right] = \sum_{i=1}^{d}\mathbf{E}[|S_i|] = \sum_{i=1}^{d}1p_i + 0(1-p_i) = \sum_{i=1}^{d}p_i.$$

E42

If $\mathbf{E}[\mathbf{C}^\top\mathbf{C}]$ is finite, then $\forall x \neq 0$:

$$x^T\mathbf{E}[\mathbf{C}^\top\mathbf{C}]x \geq 0$$
$$\mathbf{E}[x^T\mathbf{C}^\top\mathbf{C}x] \geq 0$$
$$x^T\mathbf{C}^\top\mathbf{C}x \geq 0$$
$$(\mathbf{C}x)^\top(\mathbf{C}x) \geq 0.$$

[P11]

E47

Define base case:

$$C_{1,2} := C_1 \circ C_2 \in \mathbb{B}^d(\underbrace{(\omega_1 + 1)(\omega_2 + 1) - 1}_{\omega_{1,2}}),$$

$$C_{1,3} := C_1 \circ C_2 \circ C_3 = C_{1,2} \circ C_3,$$
$$\omega_{1,3} := (\omega_{1,2} + 1)(\omega_3 + 1) - 1$$
$$= ((\omega_1 + 1)(\omega_2 + 1) - 1 + 1)(\omega_3 + 1) - 1$$
$$= (\omega_1 + 1)(\omega_2 + 1)(\omega_3 + 1) - 1,$$
$$C_{1,n} := C_1 \circ C_2 \circ ... \circ C_n = C_{1,n-1} \circ C_n,$$
$$\omega_{1,n} := (\omega_1 + 1)(\omega_2 + 1)...(\omega_n + 1) - 1 = (\omega_{1,n-1} + 1)(\omega_n + 1) - 1.$$

By induction, the base case is clear. Next, if $n = k$, assume

$$C_{1,k} := C_1 \circ C_2 \circ ... \circ C_k = C_{1,k-1} \circ C_k,$$
$$\omega_{1,k} := (\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1) - 1 = (\omega_{1,k-1} + 1)(\omega_k + 1) - 1$$

is true. Then for $n = k + 1$:

$$C_{1,k+1} := C_1 \circ C_2 \circ ... \circ C_k \circ C_{k+1} = C_{1,k-1} \circ C_k \circ C_{k+1},$$
$$\omega_{1,k+1} := (\omega_{1,k} + 1)(\omega_{k+1} + 1) - 1 = (\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1)(\omega_{k+1} + 1) - 1$$
$$:= ((\omega_{1,k-1} + 1)(\omega_k + 1) - 1 + 1)(\omega_{k+1} + 1) - 1 = (\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1)(\omega_{k+1} + 1) - 1$$
$$:= ((\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1))(\omega_{k+1} + 1) - 1 = (\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1)(\omega_{k+1} + 1) - 1,$$
$$\omega_{1,k+1} = (\omega_1 + 1)(\omega_2 + 1)...(\omega_k + 1)(\omega_{k+1} + 1) - 1.$$

E48

Define

$$\min\{a_i, b_i\} = \begin{cases} a_i, & \text{if } a_i < b_i \\ b_i, & \text{if } a_i > b_i \end{cases}.$$

Thus

$$\sum_i \min\{a_i, b_i\} = \begin{cases} \sum_i a_i, & \text{if } a_i < b_i, \forall i \\ \sum_i b_i, & \text{if } a_i > b_i, \forall i \end{cases}.$$

In case of inequality, define

$$I := \{i | a_i < b_i\},$$
$$J := \{i | a_i > b_i\}.$$

Thus

$$\sum_i \min\{a_i, b_i\} < \begin{cases} \sum_i a_i, \text{ if } |I| > |J| \\ \sum_i b_i, \text{ if } |J| > |I| \end{cases}.$$

[P12]

E55

The DCGD-SHIFT has the same exact steps in the algorithm as DCGD ($n \geq 1$ case), the difference is the gradient estimator:

$$g_h(x) := \frac{1}{n}\sum_{i=1}^n g_{h_i}(x) = \frac{1}{n}\sum_{i=1}^n h_i + \mathcal{C}_i(\nabla f_i(x) - h_i). \quad (16)$$

which means the gradients on the workers are shifted and then compressed. In order to prove the convergence theorem, first decompose

$$\mathbf{E}\left[||g_h(x^k) - \nabla f(x^\star)||^2\right] = \mathbf{E}\left[||g_h(x^k) - \nabla f(x^k)||^2\right] + ||\nabla f(x^k) - \nabla f(x^\star)||^2.$$

Then, bound

$$\mathbf{E}\left[||g_h(x^k) - \nabla f(x^k)||^2\right] = \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \underbrace{\mathcal{C}_i(\nabla f_i(x^k) - h_i) + h_i - \nabla f_i(x^k)}_{b_i^k}\right\|^2\right]$$

$$= \frac{1}{n^2}\mathbf{E}\left[\sum_i ||b_i^k||^2 \sum_{i\neq j}\langle b_i^k, b_j^k\rangle\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[||b_i^k||^2\right] + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\langle \mathbf{E}[b_i^k], \mathbf{E}[b_j^k]\rangle}_{0}$$

$$= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left\|\mathcal{C}_i(\nabla f_i(x^k) - h_i) + h_i - \nabla f_i(x^k)\right\|^2\right]$$

$$\leq \frac{1}{n^2}\sum_{i=1}^n \omega_i||\nabla f_i(x^k) - h_i||^2$$

$$= \frac{1}{n^2}\sum_{i=1}^n \omega_i||\nabla f_i(x^k) - \nabla f_i(x^\star) - (h_i - \nabla f_i(x^\star))||^2$$

$$\leq \frac{2}{n^2}\sum_{i=1}^n \omega_i||\nabla f_i(x^k) - \nabla f_i(x^\star)||^2 + \omega_i||h_i - \nabla f_i(x^\star)||^2$$

$$\leq \frac{2}{n^2}\sum_{i=1}^n 2\omega_i L_i D_{f_i}(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i||h_i - \nabla f_i(x^\star)||^2$$

$$\leq \frac{4}{n}\max(L_i\omega_i)\frac{1}{n}\sum_{i=1}^n D_{f_i}(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i||h_i - \nabla f_i(x^\star)||^2$$

$$\leq \frac{4}{n}\max(L_i\omega_i)D_{f_i}(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i||h_i - \nabla f_i(x^\star)||^2.$$

8

Thus

$$\mathbf{E}\left[||g_h(x^k) - \nabla f(x^\star)||^2\right] \leq 2\underbrace{\left(L + \frac{2}{n}\max(\omega_i L_i)\right)}_{A} D_f(x^k, x^\star) + \underbrace{\frac{2}{n^2}\sum_{i=1}^{n}\omega_i||h_i - \nabla f_i(x^\star)||^2}_{C}.$$

[P13]

E56

Thm 94: whenever $B = 0$ and $M = 0$, then $\frac{B + M\tilde{B}}{M} = 0$, proof:
First, the stepsize γ satisfies

$$0 < \gamma < \frac{1}{\mu}. \tag{17}$$

Then the iterates $\{x^k, \sigma^k\}$ satisfy

$$\mathbf{E}[d^k] \leq (1 - \gamma\mu)^k d^0 + \frac{C\gamma}{\mu}. \tag{18}$$

where

$$d^k := \left\|x^k - x^\star\right\|^2. \tag{19}$$

From Lemma 95, it is clear

$$\mathbf{E}[d^{k+1}] \leq (1 - \gamma\mu)\mathbf{E}[d^k] + C\gamma^2.$$

By recurrence, we obtain

$$\mathbf{E}[d^k] \leq (1 - \gamma\mu)^k d^0 + \frac{C\gamma}{\mu}.$$

[P14]

E57

In case of arbitrary p, the gradient estimator of L-SVRG is

$$g^k := g(x^k) - g(y^k) + \nabla f(y^k).$$

Hence, the unbiasedness:

$$\begin{aligned}
\mathbf{E}[g^k|x^k, y^k] &= \mathbf{E}[g(x^k) - g(y^k) + \nabla f(y^k)|x^k, y^k] \\
&= \mathbf{E}[g(x^k)|x^k, y^k] - \mathbf{E}[g(y^k)|x^k, y^k] + \mathbf{E}[\nabla f(y^k)|x^k, y^k] \\
&= \nabla f(x^k) - \nabla f(y^k) + \nabla f(y^k) \\
&= \nabla f(x^k).
\end{aligned}$$

E58

If
$$g(x) = \nabla f(x) + \xi,$$

then
$$g^k = (\nabla f(x^k) + \xi) - (\nabla f(y^k) + \xi) + \nabla f(y^k)$$
$$= \nabla f(x^k),$$

which is exactly GD's gradient estimator, where in this case p does not have any role, since the gradient estimator does not depend on y^k anymore. The convergence rate in this case, with stepsize $\gamma = \frac{1}{6A''}$ is

$$\mathbf{E}[d^k] \le \left(1 - \frac{\mu}{6A''}\right)^k d^0,$$

where
$$d^k := \left\| x^k - x^\star \right\|^2.$$

Thus
$$k \ge \frac{6A''}{\mu} \log \frac{1}{\epsilon},$$

which is equal to GD's rate of $\mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$.

[P15]

E62

(200) holds since the algorithm is equivalent to CGD if we set

$$x^k := h_i^k,$$
$$g^k := \mathcal{C}_i^k(\nabla f_i(x^k) - h_i^k).$$

Corollary 49 says that, if $0 < \gamma \le \frac{1}{(\omega+1)L}$ and $\nabla f(y) = 0$, then

$$\mathbf{E}\left[\left\| x^k - x^\star \right\|^2\right] \le (1 - \gamma\mu)^k \left\| x^0 - x^\star \right\|^2.$$

E63

The update rule
$$h^{k+1} = h^k - \alpha\mathcal{C}(h^k - \nabla f(x^k))$$

can be interpreted as a *descent* rule instead of *ascent*, and minimize instead of maximize.