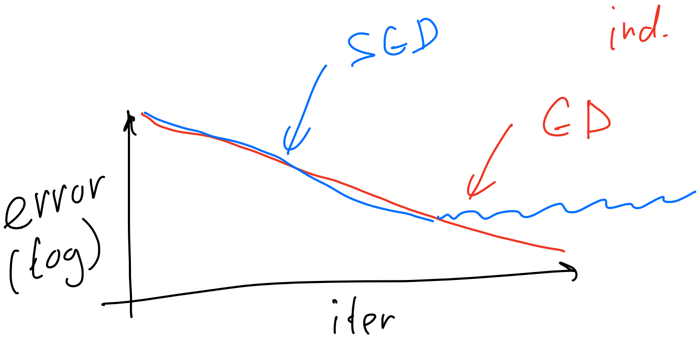
SGD:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n} L(g(x, a_i); b_i)$$

$$x^{k+1} = x^k - \gamma \nabla f_{i_k}(x^k)$$

ind. unif.



GD: $\nabla f(x) \to \nabla f(x^*) \to 0$

SGD: $\nabla f_{i_k}(x) \to \nabla f_{i_k}(x^*) \not\to 0$

$$\frac{\nabla f(x) = 0}{\boxed{\nabla f_i(x) \neq 0}}$$

$$x^{k+1} = x^k - \gamma g^k$$

what to use?

$g^k \to 0$

SAGA:

SGD 1 vector

SAGA n vectors memory

$$y_i^0 = \nabla f_i(x^0)$$

$k \neq 0$ $$y_i^k = \begin{cases} y_i^{k-1} & i \neq i_k \\ \nabla f_i(x^k) & i = i_k \end{cases}$$

$|S_k| = b$

indep't unif $[1 \dots n]$

$$g^k = \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} + \left(\nabla f_{i_k}(x^k) - y_{i_k}^{k-1}\right)$$

$\frac{1}{n}$ with mod.

modification to make $\mathbb{E}_k[g^k] = \nabla f(x^k)$

$$\mathbb{E}\|x^{k+1} - x^*\|_2^2 = \mathbb{E}\|x^k - \gamma g^k - x^*\|_2^2$$
$$= \mathbb{E}\left[\|x^k - x^*\|_2^2\right] - 2\gamma \mathbb{E}\left[\langle g^k; x^k - x^* \rangle\right] + \gamma^2 \mathbb{E}\left[\|g^k\|_2^2\right]$$

$$\mathbb{E}[\ \ ] = \mathbb{E}\left[\underbrace{\mathbb{E}_k}_{\text{cond. (take into account only } k \text{ iteration)}}[\ \ ]\right]$$

$$\mathbb{E}\left[\langle g^k; x^k - x^* \rangle\right] = \mathbb{E}\left[\mathbb{E}_k\left[\langle g^k; \underbrace{x^k - x^*}_{\text{is not stoch}} \rangle\right]\right]$$
$$= \mathbb{E}\left[\langle \mathbb{E}_k[g^k]; x^k - x^* \rangle\right]$$

SGD: $g^k = \nabla f_{i_k}(x^k)$

$$\mathbb{E}_k[g^k] = \sum_{i=1}^n \mathbb{P}\{i_k = i\} \cdot \nabla f_i(x^k)$$
$$= \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^k) = \nabla f(x^k)$$

SAGA: $\mathbb{E}_k[g^k] =$

$$= \mathbb{E}_k\left[\frac{1}{n}\sum y_i^k\right] =$$

$\mathbb{E}[a+b] = \mathbb{E}[a] + \mathbb{E}[b]$

correct

$$= \frac{1}{n}\sum \mathbb{E}_k[y_i^k] =$$
$$= \frac{1}{n}\sum \left(\nabla f_i(x^k) \cdot \frac{1}{n} + \left(1 - \frac{1}{n}\right) y_i^{k-1}\right)$$
$$= \frac{1}{n^2}\sum \nabla f_i(x^k) + \frac{1}{n}\left(1 - \frac{1}{n}\right)\sum y_i^{k-1}$$
$$= \frac{1}{n}\nabla f(x^k) + \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n}\sum y_i^{k-1}$$

want? $= \nabla f(x^k)$

after modif

$$\mathbb{E}_k[g^k] = \mathbb{E}_k\Big[\underbrace{\tfrac{1}{n}\sum y_i^{k-1}}_{\text{not rand}} + \underbrace{\nabla f_{i_k}(x^k) - y_{i_k}^{k-1}}\Big]$$

$$= \underbrace{\frac{1}{n}\sum y_i^{k-1}}+\underbrace{\mathbb{E}\left[\nabla f_{i_k}(x^k) - y_{i_k}^{k-1}\right]}$$

$$= \frac{1}{n}\sum y_i^{k-1} + \nabla f(x^k) - \frac{1}{n}\sum y_i^{k-1}$$

$$= \nabla f(x^k) \quad +$$

$$\mathbb{E}_k\left[\|g^k\|_2^2\right] = \mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|_2^2\right]$$

$$\mathbb{E}_k[\|g^k - \nabla f(x^k)\|_2^2$$

$$= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n y_i^{k-1} + \nabla f_{i_k}(x^k) - y_{i_k}^{k-1} - \nabla f(x^*)\right\|_2^2\right]$$

rand.

$$= \frac{1}{n}\sum_{j=1}^n \left\|\frac{1}{n}\sum_{i=1}^n y_i^{k-1} + \nabla f_j(x^k) - y_j^{k-1} - \nabla f(x^*)\right\|_2^2$$

$=$

$$g^k = \frac{1}{n}\sum_{i=1}^n y_i^k$$

1st

$$= \boxed{\frac{1}{n}\sum_{i\neq i_k} y_i^{k-1}} + \frac{1}{n}\left(\nabla f_{i_k}(x^k)\right)$$

$$= \boxed{\frac{1}{n}\sum_{i=1}^{n} y_i^{k-1}} - \cancel{\frac{1}{n}}\left(y_{i_k}^{k-1} - \nabla f_{i_k}(x^k)\right)$$

in new version to get $\mathbb{E}_k[g^k] = \nabla f(x^k)$

$$= \frac{1}{n}\sum_{j=1}^{n} \left\| \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} + \nabla f_j(x^k) - y_j^{k-1} - \nabla f(x^*) \right\|_2^2$$

$$= \frac{1}{n}\sum_{j=1}^{n} \left\| \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} + \nabla f_j(x^k) - y_j^{k-1} - \frac{1}{n}\sum \nabla f_i(x^*) \right\|_2^2$$

$$= \frac{1}{n}\sum_{j=1}^{n} \Big\| \nabla f_j(x^k) \underbrace{- \nabla f_j(x^*)}_{\text{add}} \underbrace{+ \nabla f_j(x^*)}_{\text{remove}}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} - y_j^{k-1} - \frac{1}{n}\sum \nabla f_i(x^*) \Big\|_2^2$$

CS $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$

$$\leq \frac{2}{n}\sum_{j=1}^{n} \left\|\nabla f_j(x^k) - \nabla f_j(x^*)\right\|_2^2$$

$$+ \frac{2}{n}\sum_{j=1}^{n} \left\| \frac{1}{n}\sum_i y_i^{k-1} - y_j^{k-1} - \frac{1}{n}\sum_i \nabla f_i(x^*) + \nabla f_j(x^*) \right\|_2^2$$

L-smothess of f_j convex $\quad \|\nabla f(x)-\nabla f(y)\|_2^2 \le 2L\left(f(x)-f(y) - \langle \nabla f(y); x-y\rangle\right)$

$$\le \frac{4L}{n}\sum_{j=1}^{n}\left(f_j(x^k) - f_j(x^*) - \langle \nabla f_j(x^*); x^k - x^*\rangle\right)$$

$$+ \frac{2}{n}\sum_{j=1}^{n}\left\| \frac{1}{n}\sum_i y_i^{k-1} - y_j^{k-1} - \frac{1}{n}\sum_i \nabla f_i(x^*) + \nabla f_j(x^*)\right\|_2^2$$

$$= 4L\left(f(x^k) - f(x^*)\right)$$

$$+ \underbrace{\frac{2}{n}\sum_{j=1}^{n}\left\| \frac{1}{n}\sum_i y_i^{k-1} - \frac{1}{n}\sum_i \nabla f_i(x^*) - y_j^{k-1} + \nabla f_j(x^*)\right\|_2^2}$$

$$\mathbb{E}\left[\|\underbrace{\xi}_{r.v} - \underbrace{\mathbb{E}\xi}_{exp.}\|_2^2\right] \overset{?}{\le} \mathbb{E}\left[\|\xi\|_2^2\right]$$

$$\underbrace{2\,\mathbb{E}_k\left[\left\| \underbrace{\frac{1}{n}\sum_i y_i^{k-1} - \frac{1}{n}\sum_i \nabla f_i(x^*)}_{\mathbb{E}_k} - \underbrace{y_{i_k}^{k-1} + \nabla f_{i_k}(x^*)}\right\|_2^2\right]}$$

$$\le 2\,\mathbb{E}_k \left\|\nabla f_{i_k}(x^*) - y_{i_k}^{k-1}\right\|_2^2$$

$$= 2\cdot\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^*) - y_i^{k-1}\right\|_2^2$$

$$\mathbb{E}\|x^{k+1} - x^*\|_2^2 = \mathbb{E}\left[\|x^k - x^*\|_2^2\right] - 2\gamma\,\mathbb{E}\left[\langle \nabla f(x^k), x^k - x^* \rangle\right]$$

$$+ \gamma^2 \cdot 4L\left(f(x^k) - f(x^*)\right)$$

$$+ \gamma^2 \cdot 2 \cdot \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - y_i^{k-1}\|_2^2$$

if $y_i^k \to \nabla f_i(x^*)$

$$\mathbb{E}_k\left[\|\nabla f_i(x^*) - y_i^k\|_2^2\right] =$$

$$= \frac{1}{n} \cdot \|\nabla f_i(x^*) - \nabla f_i(x^k)\|_2^2 + \left(1 - \frac{1}{n}\right)\|\nabla f_i(x^*) - y_i^{k-1}\|_2^2$$

$$\mathbb{E}_k\left[\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - y_i^k\|_2^2\right] = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x^k)\|_2^2$$

$$+ \left(1 - \frac{1}{n}\right)\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - y_i^{k-1}\|_2^2$$

L-smothess and conv.

$$\leq 2L \cdot \frac{1}{n}\sum\left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^* \rangle\right)$$

$$= 2L\left(f(x^k) - f(x^*)\right)$$

$$\mathbb{E}\|x^{k+1}-x^*\|_2^2 = \mathbb{E}\left[\|x^k-x^*\|_2^2\right] - \underbrace{2\gamma\mathbb{E}\left[\langle \nabla f(x^k), x^k-x^*\rangle\right]}_{\text{str conv.}}$$
$$+\gamma^2\cdot 4L\left(f(x^k)-f(x^*)\right)$$
$$\gamma^2\cdot 2\cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)-y_i^{k-1}\|_2^2$$

$$\underbrace{\mathbb{E}_k\left[\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)-y_i^{k}\|_2^2\right]}_{\sigma_k^2} \le 2L\left(f(x^k)-f(x^*)\right) + \underbrace{\left(1-\frac{1}{n}\right)}_{<1}\cdot \underbrace{\frac{1}{n}\sum\|\nabla f_i(x^*)-y_i^{k-1}\|_2^2}_{\sigma_{k-1}^2}$$

$$\mathbb{E}_k\left[\|x^{k+1}-x^*\|_2^2 + \underbrace{M\sigma_k^2}_{>0}\right] \le (1-\mu\gamma)\|x^k-x^*\|_2^2$$
$$+\left(\gamma^2\cdot 4L-\gamma\right)\left(f(x^k)-f(x^*)\right)$$
$$+2\sigma_{k-1}^2$$
$$+2LM\left(f(x^k)-f(x^*)\right)$$
$$+\left(1-\frac{1}{n}\right)M\sigma_{k-1}^2$$

need to remove

$$M = 3n \qquad 2+3n\cdot\left(1-\frac{1}{n}\right) =$$
$$= 3n-3+2 = 3n-1 =$$
$$= 3n\left(1-\frac{1}{3n}\right)$$
$$= M\left(1-\frac{1}{3n}\right)$$

$$\|x^{k+1} - x^*\|_2^2 + M\sigma_k^2 \leq (1-\mu\gamma)\|x^k - x^*\|_2^2$$
$$\left(1 - \frac{1}{3n}\right) M\sigma_{k-1}^2$$

$$\gamma < \frac{1}{2L}$$

$$\leq \max\left(1-\mu\gamma;\ 1 - \frac{1}{3n}\right)$$
$$\left(\|x^k - x^*\|_2^2 + M\sigma_{k-1}^2\right)$$

SAGA: $\left(n + \frac{L}{\mu}\right)\log\frac{1}{\varepsilon}$ iterations

SGD sublinear

GD $\frac{L}{\mu}\log\frac{1}{\varepsilon}$ iterations

SAGA in terms of comput n times better than GD

$$x^{k+1} = x^k - \gamma g^k$$

SAGA: $g^k \to 0$ $\qquad y_i^k \to \nabla f_i(x^*) \neq 0$

$\frac{1}{n}\sum y_i^k \to 0$

SVRG: $g^k = \underline{\nabla f_{\boxed{i_k}}(x^k)} - \underline{\nabla f_{i_k}(w^k)} + \underline{\nabla f(w^k)}$

w^k - reference point

$w^k = x^k$ (sometime e.g. per epoch)

SVRG (not class): $\frac{1}{n}\left(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)\right) + \nabla f(w^k)$

without

$$= \underline{\frac{1}{n}\sum \nabla f_i(w^k)} - \underline{\frac{1}{n}\nabla f_{i_k}(w^k)} + \underline{\frac{1}{n}\nabla f_{i_k}(x^k)}$$

SARAH: $\underline{g^k = \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-1}) + g^{k-1}}$

sometimes $g^k = \nabla f(x^k)$

$$x^{k+1} = x^k - \gamma \nabla f_{i_k}(x^k)$$

$\uparrow$ batch

$\Downarrow$ other stoch.

$$x^{k+1} = x^k - \gamma [\nabla f]_{i_k}(x^k)$$

$\uparrow$ random coord.

Example

$\nabla f(x) = Ax - b \qquad [\nabla f]_{i_k} = a_{i_k}^T x - b_{i_k}$

Example $f(x)$ only z-o inf

$$\frac{f(x+\tau e) - f(x-\tau e)}{2\tau}$$

$$x^{k+1} = x^k - \gamma \cdot \underbrace{d\,[\nabla f]_{i_k}(x^k)}_{g^k}$$

uniformly and indep.

$$\mathbb{E}\left[\|x^{k+1}-x^*\|\right] = \|x^k - x^*\| - 2\gamma\langle g^k ; x^k - x^*\rangle + \gamma^2\|g^k\|_2^2$$

$$\mathbb{E}_k\left[g^k\right] = ? \qquad \mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|_2^2\right] - ?$$

$$\mathbb{E}_k\left[g^k\right] = d\begin{pmatrix} \mathbb{E}_k\left[[g^k]_i\right] \\ \vdots \\ \vdots \end{pmatrix} = d\begin{pmatrix} \vdots \\ \frac{1}{d}\left[\nabla f(x^k)\right]_i \\ \vdots \end{pmatrix} = d\cdot\frac{1}{d}\nabla f(x^k) = \nabla f(x^k)$$

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|_2^2\right] =$$

$$= \mathbb{E}_k\left[\| d\left[\nabla f(x^k)\right]_{i_k}\|_2^2\right] =$$

$$= d^2\,\mathbb{E}_k\left[\|\left[\nabla f(x^k)\right]_{i_k}\|_2^2\right]$$

$$= d^2\,\mathbb{E}_k\left[\left(\left[\nabla f(x^k)\right]_{i_k}\right)^2\right]$$

the value of i_k coord of $\nabla f(x^k)$

$$= d^2 \sum_{i=1}^{d} \frac{1}{d} \left(\nabla f(x^k)\right)_i^2 =$$

$$= d \underbrace{\sum_{i=1}^{d} \left(\nabla f(x^k)\right)_i^2}_{} = d \, \|\nabla f(x^k)\|_2^2$$

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] = \mathbb{E}\|x^k - x^*\|_2^2 - \underline{2\gamma \mathbb{E}\langle \nabla f(x^k); x^k - x^*\rangle}$$

$$+ \gamma^2 d \, \mathbb{E}\|\nabla f(x^k)\|_2^2$$

μ-str conv, L-smoothness

$$\leq \mathbb{E}\|x^k - x^*\|_2^2 - \gamma\mu \|x^k - x^*\|_2^2$$

$$- \gamma \left(f(x^k) - f(x^*)\right)$$

$$+ \gamma^2 \cdot d \cdot 2L \left(f(x^k) - f(x^*)\right)$$

$$= \underline{(1 - \gamma\mu) \|x^k - x^*\|_2^2}$$

linear conv

$$\underline{- \gamma(1 - \gamma \cdot d \cdot 2L)} \left(f(x^k) - f(x^*)\right)$$

$$1 - \gamma \cdot 2dL \geq 0$$

$$\gamma \le \frac{1}{2dL}$$

$$\mathbb{E}\,\|x^{k+1} - x^*\|_2^2 \le (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$\frac{1}{\gamma\mu} \log \frac{1}{\varepsilon}$ iterations

$\frac{dL}{\mu} \log \frac{1}{\varepsilon}$ iterations