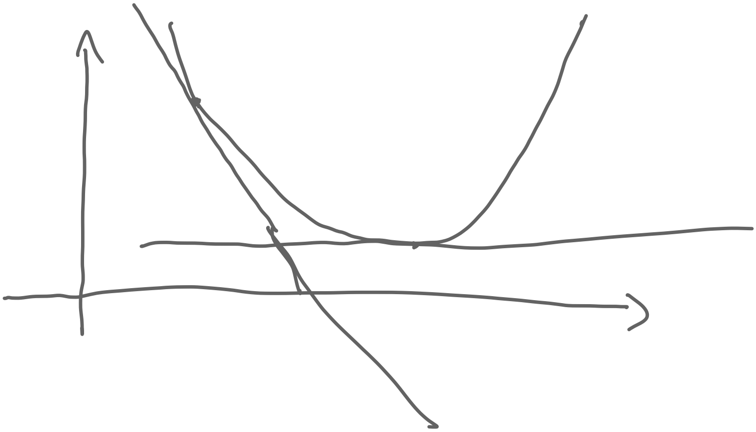
OPTIMAL LECTURE 5, 16/2/2021

- Project
- Exam



$$x_{t+1} = x_t - \eta H_F^{-1}(x_t)\, g_t$$

- CLASSIC GRADIENT METHOD
  - NOISE VARIANCE REDUCTION METHODS
- OPTIMIZATION FOR NEURAL NETWORKS

NOISE VARIANCE REDUCTION METHODS

$M \quad M_G \rightarrow$ NOISE OF GRADIENT

$$g_t = \frac{1}{n_m} \sum_{i=1}^{n_m} \underline{\nabla f(x_t, \xi_i)}$$

$X_i \sim$ i.i.d. $\quad Var(X_i) = \sigma^2$

$$Var\left(\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

$$Var\left(\nabla f(x_t, \xi)\right) = \sigma^2$$

$$Var(g_t) = \frac{\sigma^2}{n_m}$$

$$\mathbb{E}\left[F(w_{k+1}) - F^*\right] \leq \frac{\alpha L M}{2c\mu} + (1 - \alpha c\mu)^k \times$$

$$\left[F(w_1) - F_* - \frac{\alpha L M}{2c\mu}\right]$$

$$\alpha \leq \frac{\mu}{L M_G} \qquad M_G = \mu_G + M_V$$

$$\mathbb{E}\left[F(w_{k+1}) - F^*\right] \leq \frac{\alpha L M}{2c\mu} + (1-\alpha c\mu)^k \times$$

$$\left[F(w_1) - F_* - \frac{\alpha L M}{2c\mu}\right]$$

$$\alpha \leq \frac{\mu}{L M_G} \qquad M_G = \mu_G + M_V$$

Mini Batch

$$g_t = \frac{1}{n_m} \sum \nabla f(x_t, \xi_i) \qquad \mu, \mu_G$$

$$\longrightarrow \mathbb{E}[g_t]^T \nabla F(x_t) \geq \mu \|\nabla F(x_t)\|^2$$

$$\longrightarrow \|\mathbb{E}[g_t]\|^2 \leq \mu_G \|\nabla F(x_t)\|^2$$

$$\mathbb{E}[g_t] = \nabla F(w_t)$$

$$\Rightarrow \nabla F(w_t)^T \nabla F(w_t) = \|\nabla F(w_t)\|^2 \geq \mu \|\nabla F(w_t)\|^2$$

$$\Rightarrow \mu \leq 1$$

1) $= \frac{1}{|S|} \sum_{x_i \in S} \nabla f(w_t, x_i)$

2) $= \frac{1}{n_m} \sum_{i=1}^{n_m} \mathbb{E}[\nabla f(w_t, \xi_i)] = \frac{1}{n_m} n_m \nabla F(w_t) = \nabla F(w_t)$

$$\mathbb{E}\left[F(w_{k+1}) - F^*\right] \leq \frac{\alpha L M}{2c\mu} + (1-\alpha c\mu)^k \times \left[F(w_1) - F_* - \frac{\alpha L M}{2c\mu}\right]$$

$$\alpha \leq \frac{\mu}{L M_G} \qquad M_G = \mu_G + M_V$$

Mini Batch

$$g_t = \frac{1}{n_m} \sum \nabla f(x_t, \xi_i)$$

$$\mu, \mu_G$$

$$\mu = \mu_G = 1$$

$$\mathbb{V}(g_t) \leq M + M_V \|\nabla F(w_t)\|^2$$

M, M_V functions of n_m

$$M(n_m = 1) = M'$$

$$M_V(n_m = 1) = M'_V$$

$$M(n_m) = \frac{M'}{n_m} \qquad M_V(n_m) = \frac{M'_V}{n_m}$$

$$\mathbb{E}\left[F(w_{k+1}) - F^*\right] \leq \frac{\alpha L M}{2c\, n_m} + (1-\alpha c)^k \times \left[F(w_1) - F_* - \frac{\alpha L M}{2c\, n_m}\right]$$

$$\alpha \leq \frac{\mu}{L M_G}$$

$$M_G = 1 + M_V$$

MB vs SG $(n_m = 1)$

In terms of Iterations (k)
the larger n_m, the better
(the smaller the error)

What about time?
the time of 1 iteration is proportional
to n_m

In the time MB does k iterations
the SG does $n_m \times k$ iterations

MB

$$\mathbb{E}\left[F(w_{k+1}) - F^*\right] \leq \frac{\cancel{n_m}\,\alpha_{SG} L M'}{2c\,\cancel{n_m}} + \left(1 - n_m\alpha_{SG} c\right)^{k} \times$$

$$\left[F(w_1) - F_* - \frac{\cancel{n_m}\,\alpha_{SG} L M}{2c\,\cancel{n_m}}\right]$$

SG

$$\mathbb{E}\left[F(w_{n_m k+1}) - F^*\right] \leq \frac{\alpha_{SG} L M'}{2c} + \left(1 - \alpha_{SG} c\right)^{n_m k} \times$$

$$\left[F(w_1) - F_* - \frac{\alpha_{SG} L M}{2c}\right]$$

$$\alpha \leq \frac{1}{L M_G} = \frac{1}{L(1+M_V)} \simeq \frac{1}{L M_V} =$$

$$= \frac{n_m}{L M_V'}$$

$\Rightarrow$ α_{SG} I can take $\alpha_{MB} = n_m \alpha_{SG}$

MB

$$\mathbb{E}[F(w_{k+1}) - F^*] \leq \frac{\alpha_{SG} L M'}{2c} + \boxed{(1 - n_m \alpha_{SG} c)^{k} \times} \left[F(w_1) - F_* - \frac{\alpha_{SG} L M}{2c}\right]$$

SG

$$\mathbb{E}[F(w_{n_m k+1}) - F^*] \leq \frac{\alpha_{SG} L M'}{2c} + \boxed{(1 - \alpha_{SG} c)^{n_m k} \times} \left[F(w_1) - F_* - \frac{\alpha_{SG} L M}{2c}\right]$$

MB SG

$$(1 - \underline{n_m} \alpha_{SG} c)^{k} \leq (1 - \alpha_{SG} c)^{\underline{n_m} k}$$

$$n_m \alpha_{SG} c \ll 1 \qquad (1-x)^n \simeq 1 - nx$$

$$\simeq 1 - n_m \alpha_{SG} c k = 1 - \alpha_{SG} c n_m k$$

$$(1-x)^n = \sum_{i=0}^{n} \binom{n}{i} (-x)^{n-i}$$

$$\alpha \leq \frac{1}{L(1+r_v)} \simeq \frac{1}{L r_v}$$

$$r_v = \frac{r'_v}{m}$$

$$\alpha_{MB} \leq \frac{1}{L\left(1+\frac{r'_v}{n_m}\right)} = \frac{n_m}{L(n_m + r'_v)} \leq \frac{n_m}{L(1+r'_v)}$$

$$= n_m \alpha_{SG}$$

$$r'_v \gg 1 \qquad \frac{r'_v}{n_m} ? \; 1$$

$$\alpha_{MB} \leq n_m \alpha_{SG}$$

1st element : we have been optimistic for MB

2) NOT TRUE Variance decreases as $\frac{1}{n_m}$

$n_m = |S| \Rightarrow$ no variance $\sigma, \sigma_V = 0$

Variance of $\frac{1}{n} \sum_i X_i$

without resample

$$\boxed{\frac{\sigma^2}{n_m}} \underbrace{\left(1 - \frac{n_m}{|S|}\right)}$$

TOO PESSIMISTIC FOR THE MB

(MB HAS SOME ADVANTAGES WE IGNORED)

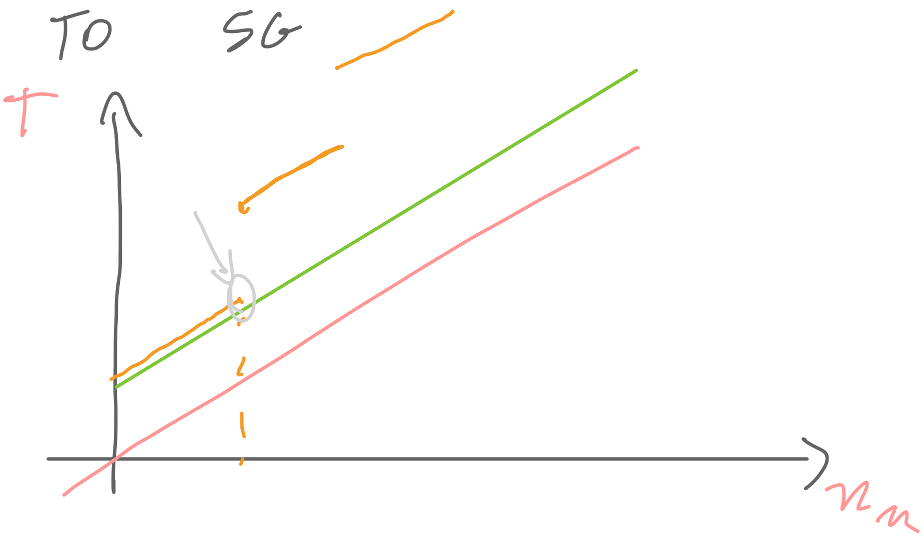
3) Probably a better computation model is

T of 1 iteration $= c + a\, n_m$

In our previous analysis we considered $c = 0$

or $a \gg c$

IT GIVES LESS ADVANTAGE TO SG



NOISE VARIANCE REDUCTION METHODS

- Basic idea
  start with small batch size and progressively increase it

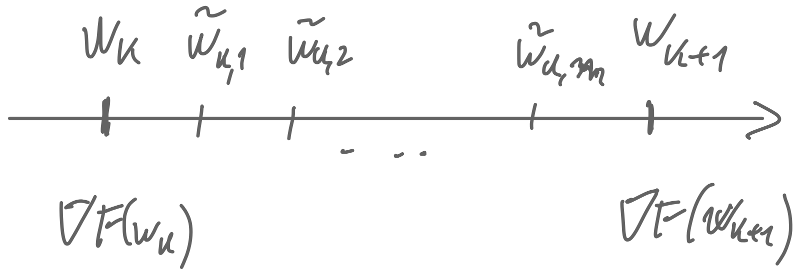
$$n_m(k) = \tau^k n_0 \qquad \tau > 1$$

YOU CAN ACHIEVE THE SAME CONVERGENCE RATE (VS TIME) OF FB

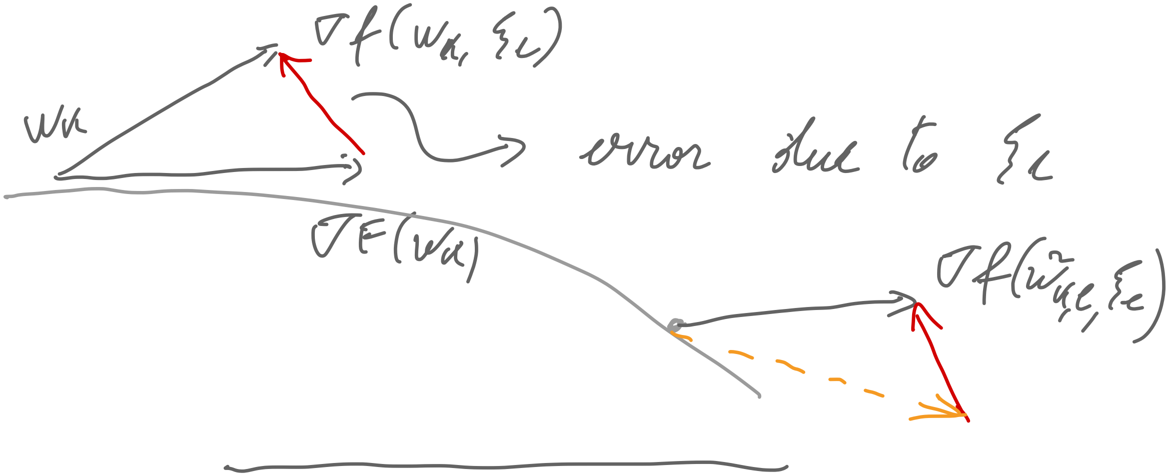
- GRADIENT AGGREGATION

SVRG SAGA

$m \in \Theta(n)$



$$\tilde{w}_{k,l+1} = \tilde{w}_{k,l} - \eta\Big(\nabla f(\tilde{w}_{k,l}, \xi_l) - \big(\nabla f(w_k, \xi_l) - \nabla F(w_k)\big)\Big)$$



SAGA

given i let $\nabla f(w_{[i]}, i)$ be the previous gradient computed on point i

$$w_{k+1} = w_k - \eta\left(\nabla f(w_k, i) - \left(\underline{\nabla f(w_{[i]}, i)} - \frac{1}{n}\sum_{j=1}^{n} \nabla f(w_{[j]}, j)\right)\right)$$

Convergence in terms of time

FB $\quad n\hat{\kappa} \ln \frac{1}{\varepsilon}$

SG $\quad \hat{\kappa}^2 \frac{1}{\varepsilon}$

$\hat{\kappa}$ condition number

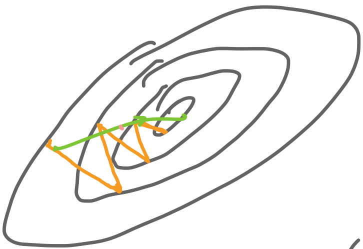
$\hat{\kappa} = \frac{L}{c}$

SAGA/SVRG $\quad (n+\hat{\kappa}) \ln \frac{1}{\varepsilon}$

SAGA, SVRG win on FB

POPULAR METHODS FOR OPTIMIZATION

- MOMENTUM



$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k) + \underbrace{\beta_k (w_k - w_{k-1})}_{\text{MOMENTUM TERM}}$$

$\alpha_k = \alpha$ $\quad$ $\beta_k = \beta$ $\quad$ SNOWBALL METHOD

FB $\quad$ $\text{error}(k) \sim \rho^k$ $\quad$ $\rho = \dfrac{\hat{K} - 1}{\hat{K} + 1}$

$\hat{K}$ condition number

Snowball $\quad$ $\text{error}(k) \sim (\rho')^k$ $\quad$ $\rho' = \dfrac{\sqrt{\hat{K}} - 1}{\sqrt{\hat{K}} + 1}$

- NESTEROV METHOD

$$w_{k+1} = w_k - \alpha \nabla f\left(\overbrace{w_k + \beta(w_k - w_{k-1})}, \xi_k\right) + \underbrace{\beta(w_k - w_{k-1})}$$

UNDER THE SAME HYPOTHESES FOR WHICH SG HAS $\frac{1}{K}$ CONV.

NESTEROV METHOD HAS $\frac{1}{K^2}$

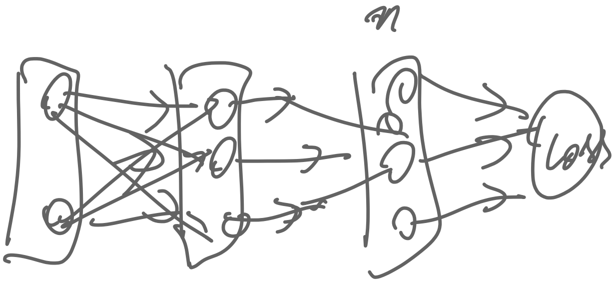
(IMPOSSIBLE TO DO BETTER)

- COORDINATE DESCENT METHOD

$$(w_{k+1})_i = (w_k)_i - \alpha \frac{\partial f(w_k, \xi)}{\partial (w_k)_i}$$

OPTIMIZATION FOR NEURAL NETWORKS

→ ADAM, ADAGRAD, RMSProp



$$L(\underline{w}) = L(\underline{o}^n) = L(o_1^n, \dots o_{m_n}^n)$$

$$o_1^n = a_1^n\left(\sum_{j=1}^{m_{n-1}} w_{j1}^{n-1} o_j^{n-1}\right)$$

output of node 1 of layer n

m_n = # of neurons of layer n

$$a_1^n(x) = a^n(x, w_1^n)$$

activation function at the 1^{st} neuron of layer n

$$l(w) = f(g(w))$$

$$l'(w) = f'\big|_{g(w)} \times g'(w)$$

$$\parallel$$

$$f'(g(w)) = f'(x)\big|_{x=g(w)}$$

$$\begin{tikzcd}
 & w_g \arrow[d] & w_f \arrow[d] & \\
x \arrow[r] & \boxed{g} \arrow[r, "g(w_g,x)"'] & \boxed{f} \arrow[r, "f(g(w_g,x), w_f)"'] & l(w)
\end{tikzcd}$$

$$l = f(w_f, \overbrace{g(w_g, x)}^{y})$$

$$\frac{\partial l}{\partial w_g} = \frac{\partial f(w_f, g(w_g, x))}{\partial g} \times \frac{\partial g(w_g, x)}{\partial w_g}$$

$$= \boxed{\left.\frac{\partial f(w_f, y)}{\partial y}\right|_{y=g(w_g,x)}} \times \frac{\partial g(w_g, x)}{\partial w_g}$$