Opt Lec 8

- Stochastic Gradient Descent II

1. Stochastic Mirror Descent (S-MD)

$\Rightarrow x_{t+1} = \underset{x \in \mathcal{X} \cap D}{\operatorname{argmin}} \ \gamma_t \tilde{g}(x_t)^T x + D_\phi(x, x_t)$ where $\mathbb{E}[\tilde{g}(x_t)] \in \partial f(x_t)$

☆ Assumptions for simple analysis

① ϕ is a mirror map that is 1-strongly convex on $\mathcal{X} \cap D$ w.r.t. $\|\cdot\|$

② $R^2 = \sup_{x \in \mathcal{X} \cap D} \phi(x) - \phi(x_1)$

2. Mini-batch SGD → reduce variance

$\Rightarrow x_{t+1} = x_t - \gamma \cdot \frac{1}{m} \sum_{i=1}^{m} g_i(x_t)$ where $g_i(x_t)$'s are i.i.d. stochastic gradients

(need more computation than SGD)

$\Rightarrow \mathbb{E}\left[\left\| \frac{1}{m} \sum_{i=1}^{m} g_i(x) - \nabla f(x) \right\|^2\right] = \mathbb{E}\left[\| g_1(x) - \nabla f(x) \|^2\right]$

3. Epoch-based algorithm (SVRG) → reduce variance

For $s = 1, 2, \cdots \infty$ → outer loop

$x_1^{(s)} = y^{(s)}$

For $t = 1, 2, \cdots, k$ → inner loop

Draw $i_t^{(s)}$ uniformly at random

$$x_{t+1}^{(s)} = x_t^{(s)} - \gamma \left(\nabla f_{i_t^{(s)}}(x_t^{(s)}) - \nabla f_{i_t^{(s)}}(y^{(s)}) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y^{(s)}) \right)$$

$y^{(s+1)} = \frac{1}{k} \sum_{t=1}^{k} x_t^{(s)}$

$\Rightarrow$

- $k \uparrow$ → computation time $\downarrow$, variance $\uparrow$
- $k \downarrow$ → computation time $\uparrow$, variance $\downarrow$

Convergence rate for stochastic mirror descent

* f is convex and $\mathbb{E}[\|\tilde{g}(x)\|_*^2] \le B^2$ $\quad (\gamma = \frac{R}{B}\sqrt{\frac{2}{T}})$

- $D_\phi(x_t, y_{t+1}) - D_\phi(x_{t+1}, y_{t+1})$

$$= \phi(x_t) - \phi(y_{t+1}) - \nabla\phi(y_{t+1})^T(x_t - y_{t+1}) - \phi(x_{t+1}) + \phi(y_{t+1}) + \nabla\phi(y_{t+1})^T(x_{t+1} - y_{t+1})$$

$$= \phi(x_t) - \phi(x_{t+1}) - \nabla\phi(y_{t+1})^T(x_t - x_{t+1})$$

$$\le \left(\nabla\phi(x_t) - \nabla\phi(y_{t+1})\right)^T(x_t - x_{t+1}) - \frac{1}{2}\|x_t - x_{t+1}\|^2$$

$$= \gamma \cdot g_t^T(x_t - x_{t+1}) - \frac{1}{2}\|x_t - x_{t+1}\|^2$$

$$\le \gamma \cdot \|g_t\|_* \|x_t - x_{t+1}\| - \frac{1}{2}\|x_t - x_{t+1}\|^2$$

$$\le \frac{\gamma^2 \|g_t\|_*^2}{2}$$

- $g_t^T(x_t - x) = \frac{1}{\gamma}\left(\nabla\phi(x_t) - \nabla\phi(y_{t+1})\right)^T(x_t - x)$

$$= \frac{1}{\gamma}\left(D_\phi(x_t, y_{t+1}) + D_\phi(x, x_t) - D_\phi(x, y_{t+1})\right)$$

$$\le \frac{1}{\gamma}\left(D_\phi(x_t, y_{t+1}) + D_\phi(x, x_t) - D_\phi(x, x_{t+1}) - D_\phi(x_{t+1}, y_{t+1})\right)$$

$$\Rightarrow \sum_{t=1}^T g_t^T(x_t - x) \le \frac{1}{\gamma}\cdot\left(D_\phi(x, x_1) - D_\phi(x, x_{T+1}) + \sum_{t=1}^T \frac{\gamma^2\|g_t\|_*^2}{2}\right)$$

$$\le \frac{R^2}{\gamma} + \frac{\gamma}{2}\sum_{t=1}^T \|g_t\|_*^2$$

- $\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T x_t\right) - f(x^*)\right] \le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T (f(x_t) - f(x^*))\right]$

$$\le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[g_t^T](x_t - x^*)\right]$$

$$\le \frac{1}{T}\left(\frac{R^2}{\gamma} + \frac{\gamma}{2}\sum_{t=1}^T \mathbb{E}[\|g_t\|_*^2]\right)$$

$$\le \frac{1}{T}\left(\frac{R^2}{\gamma} + \frac{\gamma}{2}TB^2\right)$$

$$= \frac{1}{T}\left(RB\sqrt{\frac{T}{2}} + RB\sqrt{\frac{T}{2}}\right) = RB\sqrt{\frac{2}{T}} \quad \left(\gamma = \frac{R}{B}\sqrt{\frac{2}{T}}\right)$$

(same as mirror-descent)

* f is convex, β-smooth and $\mathbb{E}[\|\nabla f(x) - \tilde{g}(x)\|_*^2] \le \sigma^2$ $\left(\gamma = \frac{1}{\beta + \frac{\sigma}{R}\sqrt{\frac{T}{2}}}\right)$

- $f(x_{t+1}) - f(x_t)$

$\le \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$

$= g_t^T (x_{t+1} - x_t) + (\nabla f(x_t) - g_t)^T (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$

$\le g_t^T (x_{t+1} - x_t) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2 + \frac{1}{2}\left(\beta + \frac{1}{\eta}\right)\|x_{t+1} - x_t\|^2$

($\because$ Cauchy-Schwarz : $2ab \le \alpha a^2 + b^2/\alpha$)

$\le g_t^T (x_{t+1} - x_t) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2 + \left(\beta + \frac{1}{\eta}\right) D_\phi(x_{t+1}, x_t)$

($\because$ 1-strongly convex : $\phi(x_{t+1}) - \phi(x_t) - \nabla\phi(x_t)^T (x_{t+1} - x_t) \ge \frac{1}{2}\|x_{t+1} - x_t\|^2$)

$= g_t^T (x^* - x_t) + g_t^T (x_{t+1} - x^*) + \left(\beta + \frac{1}{\eta}\right) D_\phi(x_{t+1}, x_t) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

$\le g_t^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(\nabla\phi(x_t) - \nabla\phi(y_{t+1})\right)^T (x_{t+1} - x^*) + \left(\beta + \frac{1}{\eta}\right) D_\phi(x_{t+1}, x_t) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

$\le g_t^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(\nabla\phi(x_t) - \nabla\phi(x_{t+1})\right)^T (x_{t+1} - x^*) + \left(\beta + \frac{1}{\eta}\right) D_\phi(x_{t+1}, x_t) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

($\because \nabla_{x_{t+1}} D(x_{t+1}, y_{t+1})^T (x_{t+1} - x^*) = \left(\nabla\phi(x_{t+1}) - \nabla\phi(y_{t+1})\right)^T (x_{t+1} - x^*) \le 0$)

$\le g_t^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})\right) + \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

($\because \left(\nabla\phi(x_t) - \nabla\phi(x_{t+1})\right)^T (x_{t+1} - x^*) = D_\phi(x^*, x_{t+1}) - D_\phi(x^*, x_t) - D_\phi(x_{t+1}, x_t)$)

$= \nabla f(x_t)^T (x^* - x_t) + (g_t - \nabla f(x_t))^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})\right)$

$+ \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

$\le f(x^*) - f(x_t) + (g_t - \nabla f(x_t))^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})\right)$

$+ \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

$\therefore f(x_{t+1}) - f(x^*) \le (g_t - \nabla f(x_t))^T (x^* - x_t) + \left(\beta + \frac{1}{\eta}\right)\left(D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})\right)$

$+ \frac{\eta}{2}\|\nabla f(x_t) - g_t\|_*^2$

$\Rightarrow \mathbb{E}[f(x_{t+1}) - f(x^*)] \le \left(\beta + \frac{1}{\eta}\right)\mathbb{E}[D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})] + \frac{\eta}{2}\sigma^2$

$\Rightarrow \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_{t+1}\right) - f(x^*)\right] \le \frac{1}{T}\left(\beta + \frac{1}{\eta}\right)\mathbb{E}[D_\phi(x^*, x_1)] + \frac{\eta}{2}\sigma^2$

$= \frac{1}{T}\left(\beta + \frac{\sigma}{R}\sqrt{\frac{T}{2}}\right)R^2 + \frac{1}{2}\cdot\frac{R}{\sigma}\sqrt{\frac{2}{T}}\sigma^2$ $\left(\eta = \frac{R}{\sigma}\sqrt{\frac{2}{T}}\right)$

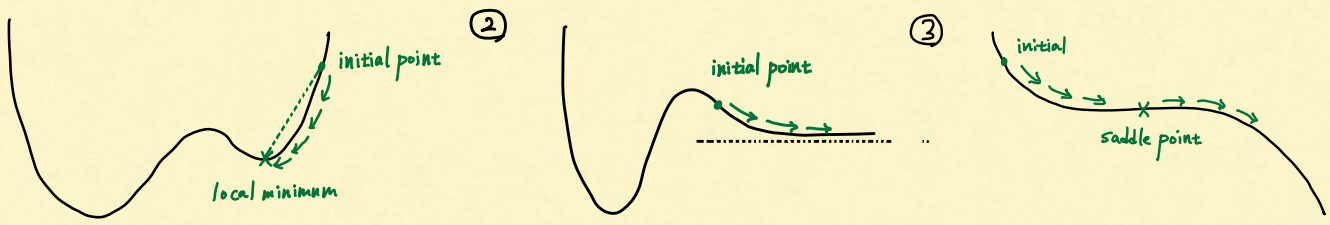
$= \frac{\beta R^2}{T} + R\sigma\sqrt{\frac{2}{T}}$

Opt Lec 9

- Non-convex optimization

1. Gradient descent on non-convex functions

① ② ③

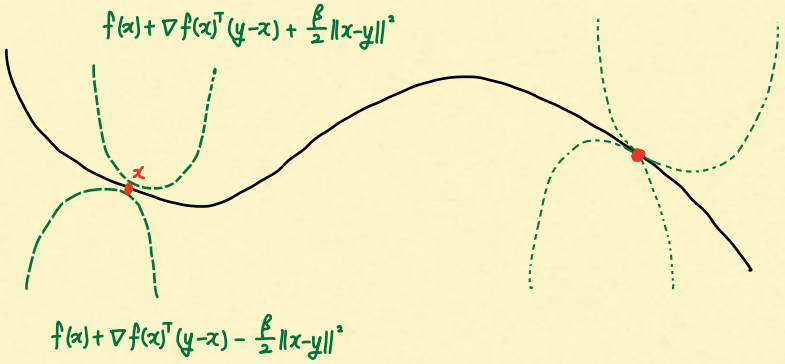


2. Smooth with parameter β

: a differentiable f satisfies follows over a convex set $\mathcal{X} \subset dom(f)$

① $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{\beta}{2}\|x-y\|^2 \quad \forall x, y \in \mathcal{X}$

② $f(y) \ge f(x) + \nabla f(x)^T (y-x) - \frac{\beta}{2}\|x-y\|^2 \quad \forall x, y \in \mathcal{X}$

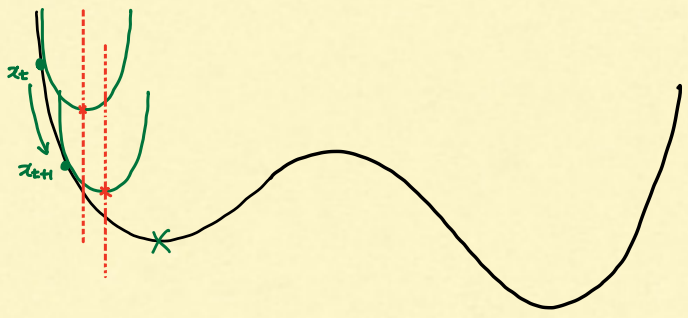


$\Rightarrow$ if f is twice differentiable, bounded Hessian ($\|\nabla^2 f(x)\| \le \beta, \forall x \in \mathcal{X}$) = smooth

($\|\cdot\|$: spectral norm $\longrightarrow$ maximum singular value of a matrix
$\longrightarrow$ maximum scale of vectors)

* No overshooting with $\gamma = \frac{1}{\beta}$

$\Rightarrow$ Never go across the local minimum to reach global minimum



Convergence rate for gradient descent

* f is twice differentiable and smooth with parameter β $\left(r = \frac{1}{\beta}\right)$

- $f(x_{t+1}) - f(x_t) \le \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \| x_{t+1} - x_t \|^2$

$$= -\frac{1}{\beta} \| \nabla f(x_t) \|^2 + \frac{1}{2\beta} \| \nabla f(x_t) \|^2$$

$$= -\frac{1}{2\beta} \| \nabla f(x_t) \|^2 \le 0 \quad \longrightarrow \text{non-increasing}$$

$$\Rightarrow \sum_{t=0}^{T-1} \| \nabla f(x_t) \|^2 \le 2\beta \left(f(x_0) - f(x_T) \right) \le 2\beta \left(f(x_0) - f(x^*) \right)$$

$$\Rightarrow \sum_{t=0}^{\infty} \| \nabla f(x_t) \|^2 \le 2\beta \left(f(x_0) - f(x^*) \right) < \infty$$

$$\therefore \lim_{t \to \infty} \| \nabla f(x_t) \|^2 = 0 \quad \Rightarrow \quad \| \nabla f(x_t) \| \longrightarrow 0$$

(but, $f(x_t) - f(x^*)$ may not converge to 0)

* Linear model

① Linear regression ($y_i = Wx_i + \varepsilon_i$)

$\Rightarrow W^* = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{n} (Wx_i - y_i)^2 \; (:= f(w)) \longrightarrow$ convex

$\longrightarrow \nabla^2 f(w) = 2 \cdot \sum_{i=1}^{n} x_i \cdot x_i^T$

$\longrightarrow \| \nabla^2 f(w) \| \le \beta \; (\ge 0)$ (largest eigenvalue)

$\longrightarrow$ gradient descent with $\gamma = \frac{1}{\beta}$

② Linear Neural Net ($y_i = W_L \cdots W_2 W_1 x_i + \varepsilon_i$)

$\Rightarrow (W_1^*, \cdots W_L^*) = \underset{W_1, \cdots W_L}{\operatorname{argmin}} \sum_{i=1}^{n} (W_L \cdots W_2 W_1 x_i - y_i)^2$

$\Downarrow \; x_i = 1, y_i = 1, n = 1$ (relaxation)

$\Rightarrow (W_1^*, \cdots W_L^*) = \underset{W_1, \cdots W_L}{\operatorname{argmin}} (W_L \cdots W_2 W_1 - 1)^2 \; (:= f(w)) \longrightarrow$ non-convex

$$\longrightarrow \nabla f(w) = 2(W_L \cdots W_2 W_1 - 1) \begin{bmatrix} W_L \cdots W_3 W_2 \\ W_L \cdots W_3 W_1 \\ \vdots \\ W_{L-1} \cdots W_2 W_1 \end{bmatrix} = 0$$

$\longrightarrow$

- $W_L \cdots W_2 W_1 = 1 \longrightarrow f(w) = 0$ (global minimum)
- At least two of W_i's are $0 \longrightarrow f(w) = 1$ (local minimum)

$\Rightarrow$ assume $W^{(0)}$ satisfies C-balanced conditions and $W_1^{(0)} W_2^{(0)} \cdots W_L^{(0)} < 1$

$\longrightarrow C$-balanced : $W_i \le C W_j \;\; \forall i \ne j$, $W_i \ge 0 \;\; \forall i$

$\longrightarrow W^{(1)} = W^{(0)} - \gamma \nabla f(W^{(0)})$ is again C-balanced.

pf) $\nabla f(W^{(0)})_i = 2(W_L \cdots W_2 W_1 - 1) \cdot \frac{W_L \cdots W_1}{W_i}$

$$\begin{aligned} (W^{(1)})_i &= (W^{(0)})_i - 2\gamma (W_L \cdots W_2 W_1 - 1) \cdot \frac{W_L \cdots W_1}{W_i} \\ &\le C \left\{ (W^{(0)})_j - 2\gamma (W_L \cdots W_2 W_1 - 1) \cdot \frac{W_L \cdots W_1}{C W_i} \right\} \\ &\le C \left\{ (W^{(0)})_j - 2\gamma (W_L \cdots W_2 W_1 - 1) \cdot \frac{W_L \cdots W_1}{W_j} \right\} \\ &= C (W^{(1)})_j \end{aligned}$$

$\therefore W^{(1)}$ is C-balanced

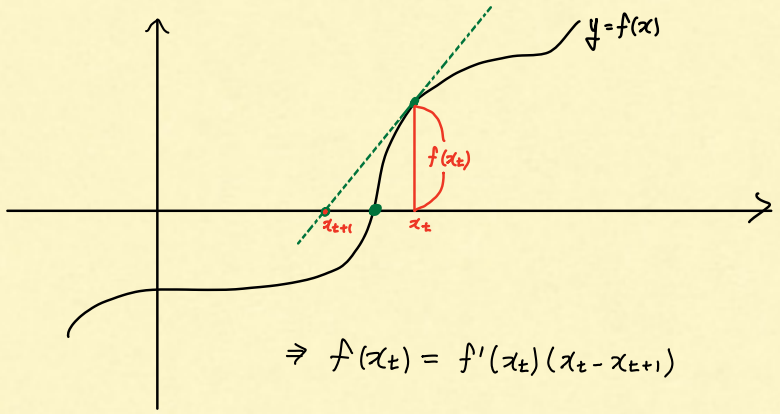
Opt Lec 10

- Newton's method

1. Newton-Raphson method

: find a zero of differentiable function f

$\Rightarrow x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$



$\Rightarrow f(x_t) = f'(x_t)(x_t - x_{t+1})$

$\Rightarrow$ Babylonian method ($f(x) = x^2 - R$)

: find a square root of R. assuming $0 < x_t - \sqrt{R} < 0.5$ and $R \geq \frac{1}{4}$

$\longrightarrow x_{t+1} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right)$

$\longrightarrow x_{t+1} - \sqrt{R} = \frac{1}{2} \cdot \frac{1}{x_t}\left(x_t^2 + R - \sqrt{R}x_t\right) = \frac{1}{2x_t}\left(x_t - \sqrt{R}\right)^2$

$\longrightarrow x_T - \sqrt{R} \leq (x_{T-1} - \sqrt{R})^2 \leq \cdots \leq (x_0 - \sqrt{R})^{2^T} \leq 2^{-2^T} \longrightarrow 0$

2. Newton's method

: find a critical point $\longrightarrow$ global minimum of a differentiable convex function f

$\Rightarrow$ control the update scale and direction according to the local geometry at x_t

$\Rightarrow$ equivalent to gradient descent when the second derivative is constant

① 1-dimensional case

$\Rightarrow$ apply Newton Rapson method to f'

$\Rightarrow x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$

② d-dimensional case

$\Rightarrow x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t)$

* Affine invariant

: independent to affine change of coordinates

$$\left[\begin{array}{l} x_t = g(y_t) = Ay_t + b \\ x_{t+1} = x_t - \nabla_{x_t}^2 f(x_t)^{-1} \nabla_{x_t} f(x_t) \\ y_{t+1} = g^{-1}(x_{t+1}) = A^{-1}(x_{t+1} - b) \end{array}\right] \Leftrightarrow \quad y_{t+1} = y_t - \nabla_{y_t}^2 f\circ g(y_t)^{-1} \nabla_{y_t} f\circ g(y_t)$$

pf) $\nabla_{y_t} f\circ g(y_t) = A^T \nabla_{x_t} f(Ay_t + b) = A^T \nabla_{x_t} f(x_t)$

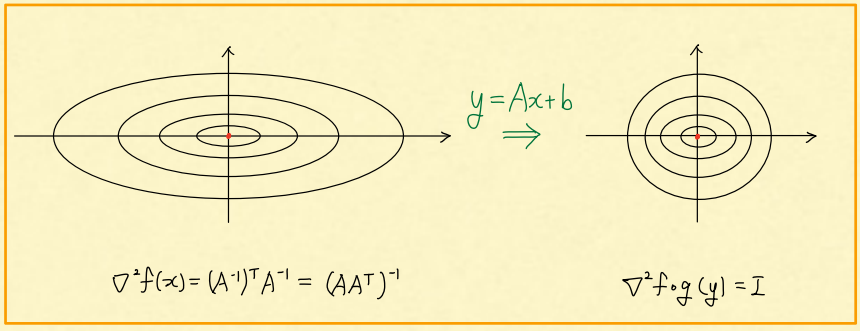
$\nabla_{y_t}^2 f\circ g(y_t) = A^T \nabla_{x_t}^2 f(Ay_t + b) A = A^T \nabla_{x_t}^2 f(x_t) A$

$y_{t+1} = y_t - \nabla_{y_t}^2 f\circ g(y_t)^{-1} \nabla_{y_t} f\circ g(y_t)$

$\Rightarrow y_{t+1} = y_t - A^{-1} \nabla_{x_t}^2 f(x_t)^{-1} \cancel{A^{-T}} \cdot \cancel{A^T} \nabla_{x_t} f(x_t)$

$\Rightarrow Ay_{t+1} + b = Ay_t + b - \nabla_{x_t}^2 f(x_t)^{-1} \nabla_{x_t} f(x_t)$

$\Rightarrow x_{t+1} = x_t - \nabla_{x_t} f(x_t)^{-1} \nabla f(x_t)$



* Another interpretation

$$\Rightarrow x_{t+1} = \operatorname{argmin}_x \left(f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2} (x - x_t)^T \nabla^2 f(x_t) (x - x_t) \right)$$

$$\longrightarrow \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t) = 0$$

$$\longrightarrow x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t)$$

$\Rightarrow$ One-shot algorithm when $\nabla^2 f(x) = H \quad \forall x$

$\Rightarrow$ Super fast when $\nabla^2 f(x)$ is small ($\approx$ smooth)

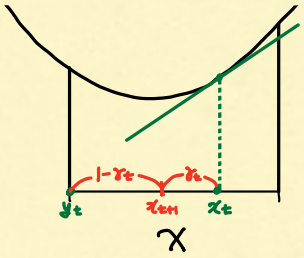
Opt Lec 11

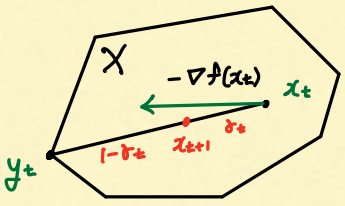
- Frank-Wolfe

1. Frank-Wolfe Algorithm

① $y_t \in \arg\min_{y \in \mathcal{X}} \nabla f(x_t)^T y$ → boundary point

② $x_{t+1} = (1-\gamma_t) x_t + \gamma_t y_t$





2. Lasso regression

$\Rightarrow \min_x \|Ax-b\|^2 \quad s.t. \quad \|x\|_1 \le 1$

$\Rightarrow y_t = -\text{sign}(\nabla f(x_t)_i)\, e_i \quad \text{where} \quad i = \text{argmax}_i \, |\nabla f(x)_i|$

$\Rightarrow x_{t+1} = (1-\gamma_t) x_t + \gamma_t y_t$

(simpler than projection on L1-ball.)

3. Duality gap ($g(x)$)

$\Rightarrow g(x_t) := \nabla f(x_t)^T (x_t - y_t) = \max_{y \in \mathcal{X}} \nabla f(x_t)^T (x_t - y)$

$\ge \nabla f(x_t)^T (x_t - x^*)$

$\ge f(x_t) - f(x^*) \ge 0 \quad (\because \text{convex})$

$\Rightarrow$ duality gap is non-negative

4. Step size

① $\gamma_t = \frac{2}{t+1}$

② line search : $\gamma_t = \arg\min_{\gamma \in [0,1]} f((1-\gamma_t) x_t + \gamma_t y_t)$

③ Gap-based : $\gamma_t = \min\left\{ \frac{g(x_t)}{\beta \|y_t - x_t\|^2}, 1 \right\}$ when f is β-smooth

Convergence rate for Frank-Wolfe

* f is convex, β-smooth and $R = \sup_{x,y \in \mathcal{X}} \|x-y\|$ $\left(\gamma_t = \frac{2}{t+1} \text{ for } t \geq 1\right)$

$\rightarrow$ For $T \geq 2$, $f(x_T) - f(x^*) \leq \frac{2\beta R^2}{T+1}$

pf) $f(x_{t+1}) - f(x_t)$

$\leq \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \|x_{t+1} - x_t\|^2$ $(\because \beta\text{-smooth})$

$= \nabla f(x_t)^T (-\gamma_t x_t + \gamma_t y_t) + \frac{\beta}{2} \|-\gamma_t x_t + \gamma_t y_t\|^2$ $(\because x_{t+1} = (1-\gamma_t)x_t + \gamma_t y_t)$

$\leq \gamma_t \nabla f(x_t)^T (y_t - x_t) + \frac{\beta}{2} \gamma_t^2 R^2$ $(\because R^2 = \sup_{x,y \in \mathcal{X}} \|x-y\|)$

$\leq \gamma_t \nabla f(x_t)^T (x^* - x_t) + \frac{\beta}{2} \gamma_t^2 R^2$ $(\because y_t = \arg\min_y \nabla f(x_t)^T y)$

$\leq \gamma_t (f(x^*) - f(x_t)) + \frac{\beta}{2} \gamma_t^2 R^2$ $(\because \text{convex})$

$\therefore f(x_{t+1}) - f(x^*) \leq (1-\gamma_t)(f(x_t) - f(x^*)) + \frac{\beta}{2} \gamma_t^2 R^2$

* Mathematical induction

① $T = 2$

$\rightarrow f(x_2) - f(x^*) \leq \frac{\beta}{2} R^2 \leq \frac{2}{3} R^2$

② $T = k$

$\rightarrow f(x_k) - f(x^*) \leq \frac{2\beta R^2}{k+1}$

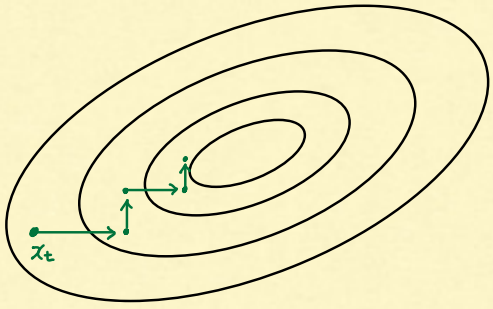
③ $T = k+1$

$\rightarrow f(x_{k+1}) - f(x^*) \leq \frac{k-1}{k+1}(f(x_k) - f(x^*)) + \frac{2\beta}{(k+1)^2} R^2$

$\leq \frac{k}{k+1} \cdot \frac{2\beta R^2}{k+1} \leq \frac{2\beta R^2}{k+2}$

Opt Lec 12

- Coordinate descent



1. Coordinate descent algorithm

: modify only one coordinate at a time

① Select $i \in [d]$

② $x_{t+1} = x_t - \gamma \boxed{\nabla_{i_t} f(x_t)} e_{i_t}$ where $\nabla_j f(x)$ indicates $\frac{\partial f(x)}{\partial x_j}$ (boxed: scalar)

$\Rightarrow$ random coordinate descent

: coordinate index is selected via uniform distribution $P(i_t = i) = \frac{1}{d} \quad \forall i$

$\rightarrow \mathbb{E}[\nabla_{i_t} f(x_t) e_{i_t}] = \frac{1}{d} \nabla f(x_t)$ $\left(\because \sum_i \nabla_i f(x_t) e_i = \nabla f(x_t)\right)$

2. Coordinate-wise β-smooth

: $|\nabla_i f(x+\gamma e_i) - \nabla_i f(x)| \le \beta |\gamma| \quad \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}, \forall i \in [d]$

$\Rightarrow f(x+\gamma e_i) \le f(x) + \gamma \nabla_i f(x) + \frac{\beta}{2}\gamma^2 \quad \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}, \forall i \in [d]$

$\Rightarrow |(\nabla^2 f)_{ii}| \le \beta$ (necessary condition for β-smooth)

3. Coordinate-wise strongly convex with α

: $|\nabla_i f(x+\gamma e_i) - \nabla_i f(x)| \ge \alpha |\gamma| \quad \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}, \forall i \in [d]$

$\Rightarrow f(x+\gamma e_i) \ge f(x) + \gamma \cdot \nabla_i f(x) + \frac{\alpha}{2}\gamma^2 \quad \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}, \forall i \in [d]$

$\Rightarrow |(\nabla^2 f)_{ii}| \ge \alpha$ (necessary condition for α-strongly convex)

Convergence rate for coordinate descent

1) $P(i_t = i) = \frac{1}{d}$ $\Rightarrow$ uniform sampling

* f is α-strongly convex and β-smooth with $\gamma = \frac{1}{\beta}$

- $f(x_{t+1}) - f(x_t) \le \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \|x_{t+1} - x_t\|^2$

  $\le -\frac{1}{\beta} \nabla f(x_t)^T \nabla_{i_t} f(x_t) e_{i_t} + \frac{1}{2\beta} \| \nabla_{i_t} f(x_t) e_{i_t} \|^2$

  $\le -\frac{1}{\beta} |\nabla_{i_t} f(x_t)|^2 + \frac{1}{2\beta} |\nabla_{i_t} f(x_t)|^2$ $(\because \nabla_i f(x) = \nabla f(x)^T e_i)$

  $= -\frac{1}{2\beta} |\nabla_{i_t} f(x_t)|^2$

- $\mathbb{E}[f(x_{t+1}) - f(x_t)] \le -\frac{1}{2\beta d} \|\nabla f(x_t)\|^2$

  $\le -\frac{1}{d} \frac{\alpha}{\beta} (f(x_t) - f(x^*))$ $(\because$ Polyak-Lojasiewicz condition$)$

- $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta}\right) \left(f(x_{T-1}) - f(x^*)\right)$

  $\le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta}\right)^T \left(f(x_0) - f(x^*)\right)$

(SGD: $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*)) + \frac{\sigma^2}{2\alpha}$)

* f is coordinate-wise β-smooth and α-strongly convex

① $\gamma = \frac{1}{\beta}$

- $f(x_{t+1}) - f(x_t) \le \nabla_{i_t} f(x_t) e_{i_t}^T (x_{t+1} - x_t) + \frac{\beta}{2} \| x_{t+1} - x_t \|^2$

$$= -\frac{1}{\beta} \| \nabla_{i_t} f(x_t) e_{i_t} \|^2 + \frac{1}{2\beta} \| \nabla_{i_t} f(x_t) e_{i_t} \|^2$$

$$\le -\frac{1}{2\beta} |\nabla_{i_t} f(x_t)|^2$$

- $\mathbb{E}[f(x_{t+1}) - f(x_t)] \le -\frac{1}{2\beta d} \| \nabla f(x_t) \|^2$

$$\le -\frac{1}{d} \frac{\alpha}{\beta} (f(x_t) - f(x^*)) \quad (\because \text{Polyak-Lojasiewicz condition})$$

- $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta}\right) (f(x_{T-1}) - f(x^*))$

$$\le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*))$$

② $\gamma = \frac{1}{\beta_{max}}$

- $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta_{max}}\right)^T (f(x_0) - f(x^*))$

$(\because \mathbb{E}[\gamma \nabla_{i_t} f(x_t) e_{i_t}] = \frac{1}{d\beta_{max}} \nabla f(x_t))$

2) $P(i_t = i) = \frac{\beta_i}{\sum_j \beta_j}$ ⇒ importance sampling

* f is coordinate-wise β_i-smooth and α-strongly convex with $\gamma_i = \frac{1}{\beta_i}$

- $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{\alpha}{\sum_j \beta_j}\right)^T (f(x_0) - f(x^*))$ → faster.

$\left(\because \mathbb{E}[\gamma_{i_t} \nabla_{i_t} f(x_t) e_{i_t}] = \frac{1}{\cancel{\beta_i}} \frac{\cancel{\beta_i}}{\sum_j \beta_j} \nabla f(x_t) = \frac{1}{\sum_j \beta_j} \nabla f(x_t)\right)$

3) $i_t = \underset{i}{\text{argmax}} |\nabla_i f(x_t)|$ ⇒ steepest coordinate descent

* f is coordinate-wise β-smooth and α-strongly convex with $\gamma = \frac{1}{\beta}$

- $\mathbb{E}[f(x_T) - f(x^*)] \le \left(1 - \frac{1}{d} \cdot \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*))$ → uniform sampling

$\left(\because \max_i |\nabla_i f(x_t)|^2 \ge \frac{1}{d} \sum_j |\nabla_j f(x)|^2\right)$

Opt Lec 13

- Duality

1. Constrained optimization problem

⇒ Primal (p^*)

$\min_x f(x)$ ⟶ convex

s.t. $h_i(x) = 0 \quad \forall\, 1 \le i \le m$ ⟶ affine

$g_j(x) \le 0 \quad \forall\, 1 \le j \le r$ ⟶ convex

⇒ Dual (d^*)

$\max_{\mu,\lambda} L(\mu,\lambda)$

s.t. $\lambda_j \ge 0 \quad \forall\, 1 \le j \le r$

Lagrange function ($\Lambda(x,\mu,\lambda)$)

⟶ $\Lambda(x,\mu,\lambda) = f(x) + \sum_{i=1}^{m} \mu_i h_i(x) + \sum_{j=1}^{r} \lambda_j g_j(x)$

Lagrange dual function ($L(\mu,\lambda)$)

⟶ $L(\mu,\lambda) = \min_x \Lambda(x,\mu,\lambda)$

⇒ primal optimal solution is lower bounded by dual feasible solution

⟶ $p^* = \inf_x f(x) \ge \inf_x \Lambda(x,\mu,\lambda) = L(\mu,\lambda)$

⇒ duality gap and stopping criterion can be defined as follow

⟶ $p^* - d^* \le p^* - L(\mu^{(k)},\lambda^{(k)}) \le f(x^{(k)}) - L(\mu^{(k)},\lambda^{(k)}) \le \varepsilon$

duality gap

stopping criterion

2. Conjugate function (f^*)

: $f^*(y) = \sup_{x \in dom(f)} y^T x - f(x)$

$\Rightarrow$ f^* is always convex as it is a pointwise supremum of affine function of y

$\rightarrow f^*(\alpha y_1 + (1-\alpha) y_2) = \sup_{x \in dom(f)} (\alpha y_1 + (1-\alpha) y_2)^T x - f(x)$

$\leq \alpha \left(\sup_{x \in dom(f)} y_1^T x - f(x) \right) + (1-\alpha) \left(\sup_{x \in dom(f)} y_2^T x - f(x) \right)$

$= \alpha f^*(y_1) + (1-\alpha) f^*(y_2)$

* $f(x) + f^*(y) \geq x^T y \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{X}^*$ (Fenchel's inequality)

pf) $f(x) + f^*(y) \geq f(x) + \sup_z z^T y - f(z) \geq f(x) + x^T y - f(x) = x^T y$

* $f^{**} \leq f$

pf) $f^{**}(x) = \sup_y y^T x - f^*(y) \leq \sup_y f(x) = f(x) \quad \forall x \in \mathcal{X}$ ($\because$ Fenchel's inequality)

(f^{**} is a closed convex of f.)

* f is closed and convex $\Rightarrow f^{**} = f^*$

* f is closed and convex $\Rightarrow y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y) \Leftrightarrow x^T y = f(x) + f^*(y)$

pf) ($\Rightarrow$) $\boxed{y \in \partial f(x)} \Rightarrow f^*(y) = \sup_z z^T y - f(z) = x^T y - f(x)$

$\Rightarrow \boxed{f^*(y) = x^T y - f(x)} = x^T y - f^{**}(x)$ ($\because f^{**} = f$)

$\Rightarrow f^{**}(x) = x^T y - f^*(y) = \sup_z z^T x - f^*(z)$

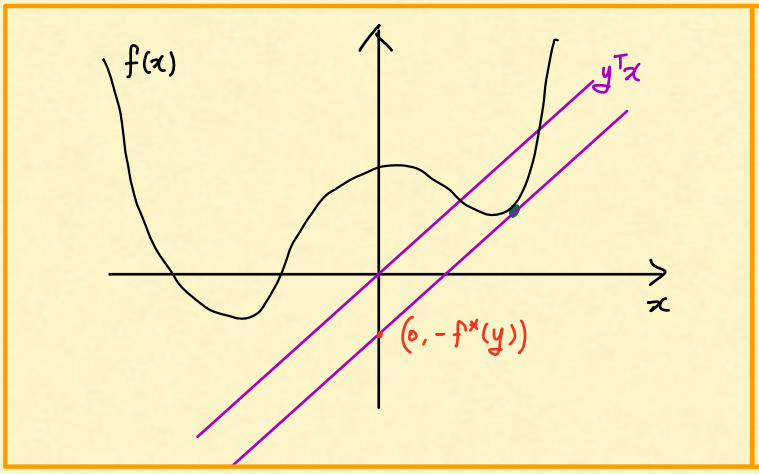
$\Rightarrow \boxed{x \in \partial f^*(y)}$

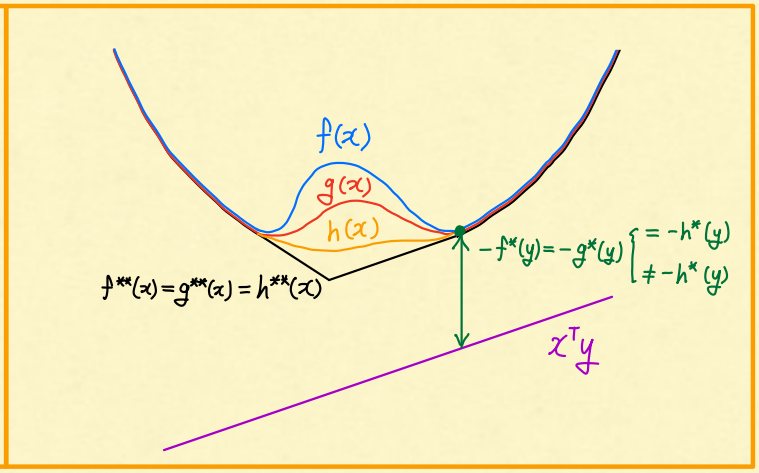
($\Leftarrow$) $\boxed{x \in \partial f^*(y)} \Rightarrow f^{**}(x) = \sup_z z^T x - f^*(z) = y^T x - f^*(y)$

$\Rightarrow f^{**}(x) = \boxed{f(x) = x^T y - f^*(y)}$ ($\because f^{**} = f$)

$\Rightarrow f^*(y) = x^T y - f(x) = \sup_z z^T y - f(z)$

$\Rightarrow \boxed{y \in \partial f(x)}$





⇒ Maximum gap between x^Ty and $f(x)$ ⇒ minimum gap between x^Ty and $f(x)$

3. Generalized linear models

⇒ primal (p^*)

$$\min_{x,w} f(w) + g(x)$$

$$\text{s.t. } w = Ax$$

⇒ dual (d^*)

$$\max_u \left(\min_{x,w} f(w) + g(x) + u^T(Ax - w) \right)$$

$$= \max_u \left(\min_{x,w} -(u^Tw - f(w)) - (-u^TAx - g(x)) \right)$$

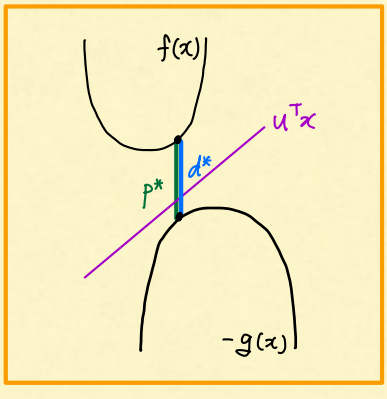
$$\geq \max_u \left(-f^*(u) - g^*(-u^TA) \right)$$

* $A = I$

$$\Rightarrow \min_x f(x) + g(x)$$

$$\Rightarrow \max_u -f^*(u) - g^*(-u)$$

$$= \max_u \left(\min \left(f(w) - u^Tw \right) + \min \left(u^Tw - (-g(w)) \right) \right)$$



* conjugate function of constant function is infinite

$$\Rightarrow g(x) = c \;\; \forall x \longrightarrow g^*(u) = \infty \;\; \forall u$$

Opt Lec 14

- Variance reduction

1. Variance reduction

: Reducing variance of X by using variable Y that is positively correlated to X with known expectation

$\Rightarrow Z_\alpha = \alpha(X-Y) + \mathbb{E}[Y]$

$\Rightarrow \mathbb{E}[Z_\alpha] = \alpha\,\mathbb{E}[X] + (1-\alpha)\,\mathbb{E}[Y]$ → unbiased when $\alpha = 1$

$\Rightarrow Var[Z_\alpha] = \alpha^2(Var(X) + Var(Y) - 2Cov(X,Y))$ → smaller variance when $\alpha < 1$

① SVRG

→ $X = \nabla f_{i_t}(\theta_t)$

→ $Y = \nabla f_{i_t}(\tilde{\theta})$, $\mathbb{E}[Y] = \frac{1}{n}\sum_i \nabla f_i(\tilde{\theta})$ → high computation

→ $\alpha = 1$ → unbiased

$\therefore x_{t+1} = x_t - \gamma\left(\nabla f_{i_t}(\theta_t) - \nabla f_{i_t}(\tilde{\theta}) + \frac{1}{n}\sum_i \nabla f_i(\tilde{\theta})\right)$

② SAG

→ $X = \nabla f_{i_t}(\theta_t)$

→ $Y = y_{i_t}^{(t-1)}$, $\mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i^{(t-1)}$ → high memory cost

$$\left(y_i^{(t)} = \begin{cases} \nabla f_i(\theta_t) & \text{if } i = i_t \\ y_i^{(t-1)} & \text{otherwise} \end{cases}\right)$$

→ $\alpha = \frac{1}{n}$ → biased

$\therefore x_{t+1} = x_t - \gamma\left(\frac{1}{n}\left(\nabla f_{i_t}(\theta_t) - y_{i_t}^{(t-1)}\right) + \frac{1}{n}\sum_i y_i^{(t-1)}\right)$

③ SAGA

→ $X = \nabla f_{i_t}(\theta_t)$

→ $Y = y_{i_t}^{(t-1)}$, $\mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i^{(t-1)}$ → high memory cost

$$\left(y_i^{(t)} = \begin{cases} \nabla f_i(\theta_t) & \text{if } i = i_t \\ y_i^{(t-1)} & \text{otherwise} \end{cases} \right)$$

→ $\alpha = 1$ → unbiased

$\therefore x_{t+1} = x_t - \gamma\left(\nabla f_{i_t}(\theta_t) - y_{i_t}^{(t-1)} + \frac{1}{n}\sum_i y_i^{(t-1)}\right)$

SVRG	SAGA	
2 Grad	1 Grad	Computation
$\tilde{\theta}, \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{\theta})$	$[y_1^t, y_2^t, \cdots, y_n^t]$	Memory storage

Opt Lec 15

- Distributed deep learning

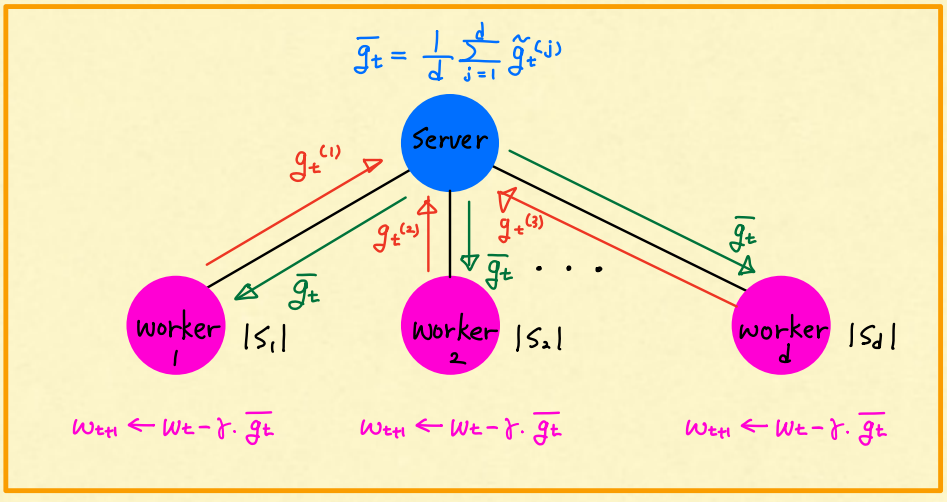
1. Distributed system

: data is distributed to separate machines as data can not fit to one device anymore

⇒ data can not be sent to outside ⟶ privacy

⇒ communication bottleneck occurs

⇒ distributed SGD with data parallelism



① Tern Grad

scalar, binary vector

$$\Rightarrow \tilde{g}_t = \text{ternalize}(g_t) = s_t \cdot \text{sign}(g_t) \circ b_t$$

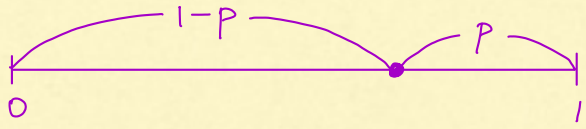
$\circ$ → elementwise product

$$\rightarrow s_t = \max(\text{abs}(g_t))$$

$$\rightarrow (b_t)_k = \begin{cases} 1 & \text{with probability } \frac{|(g_t)_k|}{s_t} \\ 0 & \text{with probability } 1 - \frac{|(g_t)_k|}{s_t} \end{cases}$$

$$\Rightarrow \mathbb{E}[(b_t)_k] = \frac{|(g_t)_k|}{s_t}$$

$$\Rightarrow \mathbb{E}[\tilde{g}_t] = g_t \quad \text{(unbiased)}$$



$$\Rightarrow \begin{cases} g_t \in \mathbb{R}^p \rightarrow 32p \\ \tilde{g}_t \in \mathbb{R}^p \rightarrow 2p+32 \end{cases}$$

② QSGD

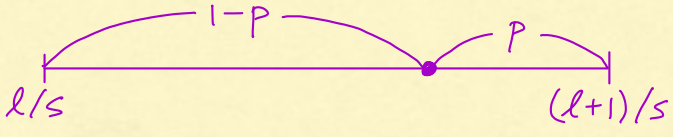
(tuning parameter ≥ 1)
# of quantization level

$$\Rightarrow (\tilde{g}_t)_k = Q_s((g_t)_k) = \|g_t\|_2 \cdot \text{sign}((g_t)_k) \cdot \xi_k(g_t, s)$$

$$\rightarrow \xi_k(g_t, s) = \begin{cases} l/s & \text{with probability } 1 - \left(\frac{|(g_t)_k|}{\|g_t\|_2} s - l\right) \\ (l+1)/s & \text{with probability } \frac{|(g_t)_k|}{\|g_t\|_2} s - l \end{cases}$$

$$\rightarrow 0 \leq l < s \text{ (integer)} \Rightarrow \frac{|(g_t)_k|}{\|g_t\|_2} \in [l/s, (l+1)/s]$$

$$\Rightarrow \mathbb{E}[\xi_k(g_t, s)] = \frac{|(g_t)_k|}{\|g_t\|_2}$$



$$\Rightarrow \mathbb{E}[(\tilde{g}_t)_k] = (g_t)_k \quad \text{(unbiased)}$$

$$\Rightarrow \text{Var}((\tilde{g}_t)_k) = \min\left(\frac{k}{s^2}, \frac{\sqrt{k}}{s}\right)\|g_t\|_2^2$$

$$\Rightarrow \begin{cases} g_t \in \mathbb{R}^p \rightarrow 32p \\ \tilde{g}_t \in \mathbb{R}^p \rightarrow 2p+32 \end{cases}$$

Opt Lec 16

- Deep learning optimization

1. SGD rules

① SGD

$$: \theta_{t+1} = \theta_t - \eta_t \nabla \ell(\theta_t)$$

② Momentum

$$: v_{t+1} = \gamma v_t + \nabla \ell(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta_t v_{t+1}$$

③ Nesterov

$$: v_{t+1} = \gamma v_t + \nabla \ell(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta_t (\gamma v_{t+1} + \nabla \ell(\theta_t))$$

④ Adagrad

$$: v_{t+1} = v_t + \nabla \ell(\theta_t)^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{\sqrt{v_{t+1} + \varepsilon}} \nabla \ell(\theta_t)$$

⑤ RMSProp

$$: v_{t+1} = \rho v_t + (1-\rho) \nabla \ell(\theta_t)^2$$

$$m_{t+1} = \gamma m_t + \frac{\eta_t}{\sqrt{v_{t+1} + \varepsilon}} \nabla \ell(\theta_t)$$

$$\theta_{t+1} = \theta_t - m_{t+1}$$

⑥ ADAM

$$: m_{t+1} = \beta_1 m_t + (1-\beta_1) \nabla \ell(\theta_t)$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2) \nabla \ell(\theta_t)^2$$

$$b_{t+1} = \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}}$$

$$\theta_{t+1} = \theta_t - \alpha_t \frac{m_{t+1}}{\sqrt{v_{t+1} + \varepsilon}} b_{t+1}$$

⑦ NADAM

$$: m_{t+1} = \beta_1 m_t + (1-\beta_1) \nabla \ell(\theta_t)$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2) \nabla \ell(\theta_t)^2$$

$$b_{t+1} = \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}}$$

$$\theta_{t+1} = \theta_t - \alpha_t \frac{\beta_1 m_{t+1} + (1-\beta_1) \nabla \ell(\theta_t)}{\sqrt{v_{t+1} + \varepsilon}} b_{t+1}$$

2. Inclusion relationship

$\Rightarrow$ SGD $\subseteq$ Momentum $(\gamma = 0)$

$\Rightarrow$ SGD $\subseteq$ Nesterov $(\gamma = 0)$

$\Rightarrow$ Momentum $\subseteq$ RMSProp $(\rho = 1, \varepsilon = 0)$

$\Rightarrow$ Momentum $\subseteq$ ADAM $(\alpha_t = \varepsilon \eta_t (1-\gamma^t), \beta_1 = \gamma, \beta_2 = 0)$

$\Rightarrow$ Nesterov $\subseteq$ NADAM $(\alpha_t = \varepsilon \eta_t (1-\gamma^t), \beta_1 = \gamma, \beta_2 = 0)$

* Complex algorithm may not perform well with bad hyperparameters

3. Learing rate & Batch size

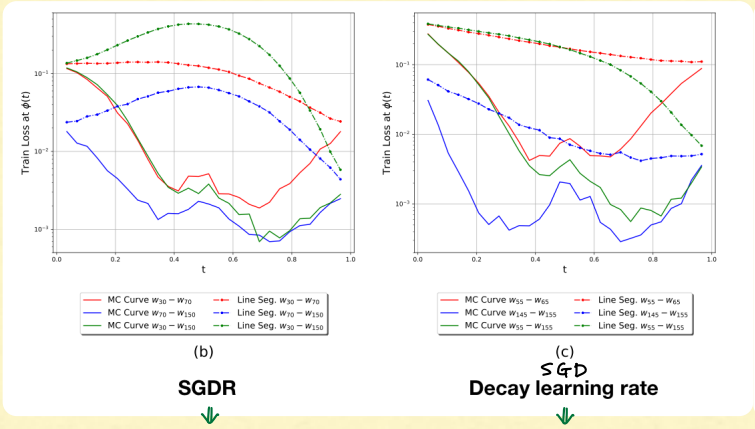
⇒ important for the following 3 cases

① objective function is non-convex

② many local minimas

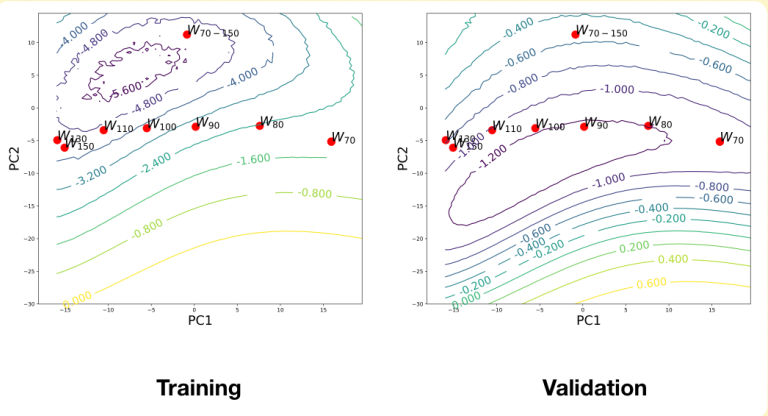
③ generalization

⇒ Learning rate



SGDR ⇒ barrier exists in line segment (o)

SGD **Decay learning rate** ⇒ barrier exists in line segment (x)



Training **Validation**

① Ensembling by mode connectivity has limitation

② Averaging iterates results in (training loss ↑ / validation loss ↓)

⇒ Batch size