Learning Theory

$\mathcal{X}$ - domain $\in \mathbb{R}^d$ $\quad D \sim \mathcal{X} \times \mathcal{Y}$

$\mathcal{Y}$ - label set $\quad \ell$ - loss f^n $\quad \mathcal{H} \times \mathcal{X} \times \mathcal{Y}$

$\mathcal{H}$ - hypothesis class $\rightarrow$ parameterized by ω
$\forall \omega$ s.t. $\|\omega\|^2 \leq B$

Goal

Output: $\underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \underset{(x,y)\sim D}{E} \left[\ell(h, x, y) \right]$

Algorithm

initialize ω_0
For $t = 1, 2, \ldots$
sample $z \sim D$
$v_t \in \partial \ell(\omega_t, z)$
$\omega_{t+1} = \omega_t - \eta_t v_t$
Output $\bar{\omega} = \frac{1}{T} \sum_{t=1} \omega^t$

Or:

$S: ((x_1, y_1), \ldots, (x_n, y_n))$

Output $\underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h, x_i, y_i)$

Algorithm: SG

initialize ω_0
For $t = 1, 2, \ldots$
sample $i_t \sim \mathcal{U}([n])$
$v_t \in \partial \ell(\omega_t, x_{i_t}, y_{i_t})$
$\omega_{t+1} = \omega_t - \eta_t v_t$
Output $\bar{\omega} = \frac{1}{T} \sum_{t=1} \omega^t$

Denote $f_i(\omega) = \ell(\omega, x_i, y_i)$

$g(\omega) = \frac{1}{n} \sum f_i(\omega)$

$g(\omega)$ is ρ-lipschitz

$\forall i, \forall u, v \in \mathbb{R}^d$
$\|f_i(u) - f_i(v)\| \leq \rho \|u - v\|$

$g(\omega)$ is L-smooth
$\forall i, \forall u, v \in \mathbb{R}^d$
$\|f_i'(u) - f_i'(v)\| \leq L \|u - v\|$

or

$\forall \omega \; |\text{eigenvalues}[g''(\omega)]| \leq L$

FG

$v_t = \frac{1}{n} \sum_{i=1}^{n} \partial \ell(\omega_t, x_i, y_i)$

$\omega_{t+1} = \omega_t - \eta_t v_t$

$g(w)$ λ-strongly convex $\qquad k = \frac{L}{\mu}$

$\forall w$ eigenvalues $[g''(w)] \geq \lambda$

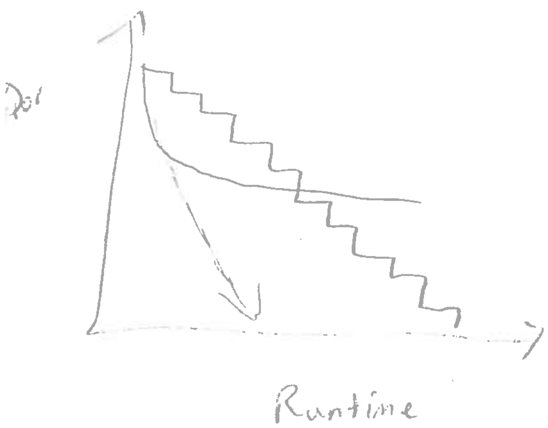
G		ϵ-error	complexity
onvex	$g(w^t) - g(w^*) \leq O(1/t)$	$\rightarrow O(1/\epsilon)$	$O(nd)$
strong onvexity	$g(w^t) - g(w^*) \leq O((1-\lambda/L)^t) = O(e^{-t\lambda/L})$	$\rightarrow O(k \log \frac{1}{\epsilon})$	

Step size backtracking line search

G

onvex	$E[g(w^t)] - g(w^*) \leq O(1/\sqrt{t})$	$\rightarrow O(1/\epsilon^2)$
strong onvexity	$E[g(w^t)] - g(w^*) \leq O(1/t)$	$\rightarrow O(1/\epsilon)$

Step size	Complexity
$\eta_t = \frac{1}{L\sqrt{t}}$	$O(d)$
$\eta_t = \frac{1}{\lambda t}$	$O(d)$



SAG (Stochastic Average Gradient)

- Store $\nabla^T f_i(\omega)$ $\forall i$
- Sample $i_t \sim \mathcal{U}([n])$
- $\omega_t = \omega_{t-1} - \eta_t \frac{1}{n} \sum y_i^t$ where $y_i^t = \begin{cases} \nabla f(\omega_{t-1}) & i = i_t \\ y_i^{t-1} & \text{o.w} \end{cases}$

initialize $sum = 0$, $y_i = 0$ $\forall i$

for $t = 1, 2 \ldots$

$i_t \sim \mathcal{U}([n])$

$sum = sum - y_{i_t} + \nabla f_{i_t}(\omega^t)$

$y_{i_t} = \nabla f_{i_t}(\omega^t)$

$\omega^{t+1} = \omega^t - \frac{\eta}{n} sum$

Thm

SAG	Step size		Complexity	Storage
convex	$\eta = \frac{1}{16L}$	$E[g(\omega^t)] - g(\omega^*) \le O(1/t)$	$O(d)$	$O(nd)$
λ-strong convexity	$\eta = \frac{1}{16L}$	$E[g(\omega^t)] - g(\omega^*) \le O\left(\left(1 - \min\left(\frac{\lambda}{16L}, \frac{1}{8n}\right)\right)^t\right)$	$O(d)$	$O(nd)$

Variance Reduction

Reduce variance of sampled X by sampling Y with known expectation

$$Z_\alpha = \alpha(X - Y) + E[Y]$$

$$E[Z_\alpha] = \alpha E[X] + (1-\alpha) E[Y]$$

$$Var(Z_\alpha) = \alpha^2 [var(X) + var(Y) - 2cov(X,Y)]$$