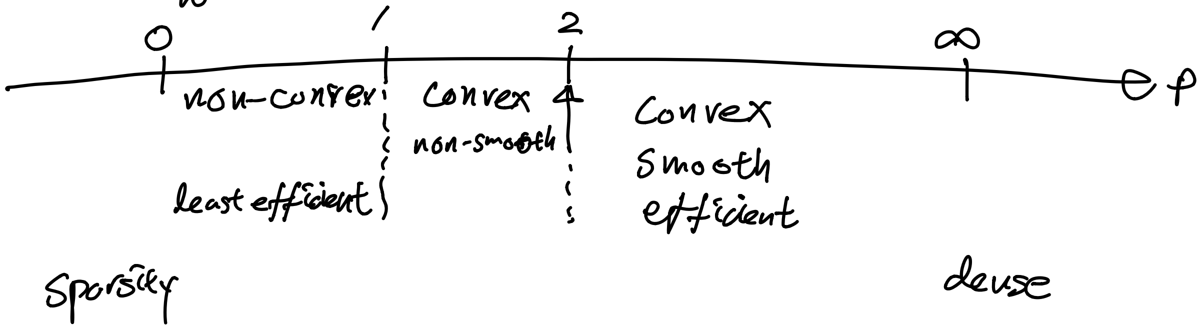
$$\hat{w} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_p^p \qquad \|w\|_p = (|w_1|^p + |w_2|^p + \dots)^{1/p}$$



Stochastic Gradient Descent

- Gradient Descent → linear regression
- Coordinate Descent → lasso
- Stochastic G.D.

Machine Learning Problems

$n \geq 1M$

$d > 1M$

- **Given data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Learning a model's parameters:** $\frac{1}{n}\sum_{i=1}^n \ell_i(w) = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T w)^2$

Gradient Descent:

$$w_{t+1} = w_t - \eta \underbrace{\nabla_w \left(\frac{1}{n}\sum_{i=1}^n \ell_i(w) \right)}_{O(d) \times n} \Big|_{w=w_t}$$

Machine Learning Problems

- **Given data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Learning a model's parameters:** $\frac{1}{n}\sum_{i=1}^n \ell_i(w) = \ell(w)$

$$\nabla \ell(w) = \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(w)$$

Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \left(\frac{1}{n} \sum_{i=1}^n \ell_i(w) \right) \Big|_{w=w_t}$$

Stochastic Gradient Descent:

old time

Random

$$w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w) \Big|_{w=w_t}$$

I_t drawn uniform at random from $\{1, \ldots, n\}$

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \sum_{i=1}^n \underbrace{\mathbb{P}(I_t = i)}_{= 1/n} \cdot \nabla \ell_i(w) = \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(w)$$

Stochastic Gradient Descent

Theorem

Let $w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w)\Big|_{w=w_t}$ I_t drawn uniform at random from $\{1,\ldots,n\}$ so that

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(w) =: \nabla \ell(w)$$

If $\|w_0 - w_*\|_2^2 \leq R$ and $\sup_w \max_i \|\nabla \ell_i(w)\|_2^2 \leq G$ then

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \leq \frac{R}{2T\eta} + \frac{\eta G}{2} \leq \sqrt{\frac{RG}{T}} = O\left(\frac{1}{\sqrt{T}}\right), \quad \eta = \sqrt{\frac{R}{GT}}$$

instead of W_T

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

(In practice use last iterate)

Stochastic Gradient Descent

show: $\mathbb{E}[\ell(w_t) - \ell(w_*)] \leq ?$, $w_{t+1} = w_t - \eta \cdot \nabla \ell_{I_t}(w_t)$

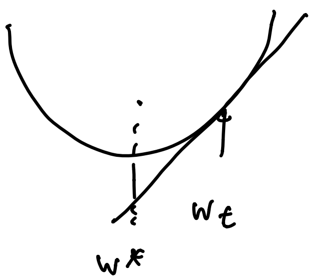
Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta \nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

$$= \mathbb{E}[\underbrace{||w_t - w_*||^2}] + \eta^2 \mathbb{E} \cdot [\underbrace{\nabla \ell_{I_t}(w_t)^2}_{\leq G}] \boxed{-2\eta \mathbb{E}[\underline{(w_t - w_*)^T \cdot \nabla \ell_{I_t}(w_t)}]}$$

$$\leq \mathbb{E}[||w_t - w_*||^2] + \eta^2 G + (\ell(w^*) - \ell(w_t)) 2\eta$$

* Convexity $\ell_i(w)$: $\ell(w^*) \geq \ell(w_t) + \nabla \ell_i(w_t)^T (w_* - w_t)$



$\rightarrow \ell(w^*) - \ell(w_t) \geq \boxed{-\nabla \ell_i(w_t)^T (w_t - w_*)}$

Stochastic Gradient Descent

Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta\nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

$$\leq \mathbb{E}[||w_t - w_*||^2] + \eta^2 G + 2\eta\, \mathbb{E}[\ell_{I_t}(w^*) - \ell_{I_t}(w_t)]$$

$$\mathbb{E}[\ell_{I_t}(w_t) - \ell_{I_t}(w^*)] \leq \frac{\mathbb{E}[||w_t - w_*||^2] - \mathbb{E}[||w_{t+1} - w_*||^2] + \eta^2 G}{2\eta}$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_{I_t}(w_t) - \ell_{I_t}(w^*)] \leq \frac{1}{2\eta T}\left\{\mathbb{E}[||w_0 - w_*||^2] - \mathbb{E}[||w_T - w_*||^2]\right\} + \cancel{T}\eta^2 G$$

Stochastic Gradient Descent

Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta \nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

$$= \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] + \eta^2 \mathbb{E}[||\nabla \ell_{I_t}(w_t)||_2^2]$$

$$\leq \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\ell(w_t) - \ell(w_*)] + \eta^2 G$$

$$\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] = \mathbb{E}\left[\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*) | I_1, w_1, \dots, I_{t-1}, w_{t-1}]\right]$$

$$= \mathbb{E}\left[\nabla \ell(w_t)^T (w_t - w_*)\right]$$

$$\geq \mathbb{E}\left[\ell(w_t) - \ell(w_*)\right]$$

$$\mathbb{E}[\ell(\bar{w})]$$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

$$\sum_{t=1}^{T} \mathbb{E}[\ell(w_t) - \ell(w_*)] \leq \frac{1}{2\eta}\left(\mathbb{E}[||w_1 - w_*||_2^2] - \mathbb{E}[||w_{T+1} - w_*||_2^2] + T\eta^2 G\right)$$

$$\leq \frac{R}{2\eta} + \frac{T\eta G}{2}$$

Stochastic Gradient Descent

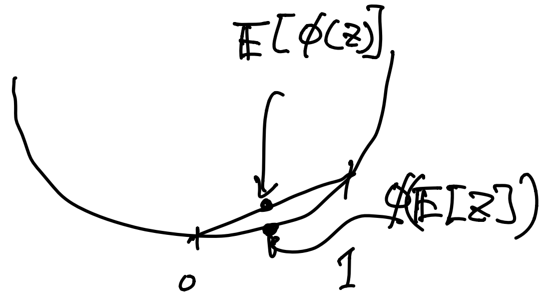
Proof

Jensen's inequality:
For any random $Z \in \mathbb{R}^d$ and convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(\mathbb{E}[Z]) \leq \mathbb{E}[\phi(Z)]$

$Z \in \{0, 1\}$

$\frac{1}{2}, \frac{1}{2}$

$\phi(z) = z^2$



$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell(w_t) - \ell(w_*)]$$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

$\{w_1 \cdots w_T\}$

u.p $\frac{1}{T} \cdots \frac{1}{T}$

$$\leq \frac{R}{2\eta T} + \frac{\eta G^2}{2}$$

$\mathbb{E}\left[\|w_0 - w_*\|^2\right]$

Stochastic Gradient Descent

$\ell(w) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(w)$ $+\|w\|_2^2$

$\|w\|_1$

(x_i, y_i)

SGD for $t = 1 \cdots T$.

$-\eta\left(\underbrace{\nabla \ell_{I_t}(w)}_{\mathbb{E}[\;] = \nabla \ell(w)} + \nabla R(w)\right)$

S.G.D.

G.D

Proof

Jensen's inequality:
For any random $Z \in \mathbb{R}^d$ and convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(\mathbb{E}[Z]) \leq \mathbb{E}[\phi(Z)]$

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell(w_t) - \ell(w_*)]$$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

$$\boxed{\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \leq \frac{R}{2T\eta} + \frac{\eta G}{2} \leq \sqrt{\frac{RG}{T}}}$$

$$\eta = \sqrt{\frac{R}{GT}}$$

η = step size

T = how many iterations.

Mini-batch SGD

Instead of one iterate, average B stochastic gradient together

$$-\eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{I_{t,j}}(w)$$

Advantages:

- Smaller variance
- Parallelization