Algorithmic ML

Lecture 5 :

Last 2 lectures :

Gradient Descent (GD/NAGD)

Analysis of GD

This lecture :

Stochastic Gradient Descent (SGD)

GD : (parameter : "step size" or "learning rate" η)

1. Choose $w_0 \in \mathbb{R}^d$
2. for $i = 1, \dots, T$:

$$w_i = w_{i-1} - \eta \nabla f(w_{i-1})$$

Thm : (GD convergence) $f : \mathbb{R}^d \to \mathbb{R}$ β-smooth & convex

$$\Rightarrow f(w_k) \leq f(w^*) + \frac{\beta \|w^* - w_0\|}{2k} \quad \text{when } \eta = 1/\beta .$$

↳ if $\beta \cdot \|x_* - x_0\| = O(1)$

then GD returns w such that $f(w) \leq f(w^*) + \varepsilon$

in $O(1/\varepsilon)$ iterations.

Pro : Always converges in $O(1/\varepsilon)$ iterations.

Con : What is the time complexity per iteration ?

→ Cost of computing $\nabla f(w)$.

Taking a step back: what f are we working with?

Recall ERM: (empirical risk minimization)

Given labelled examples ("training data") $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ ... empirical

Define $L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x_i), y_i)$ where — how bad is h_w at explaining (x_i, y_i)

Empirical loss function

Our "f" in GD. Trying to find w^* that minimizes this loss.

- $h_w : \mathbb{R}^d \to \mathbb{R}$ the hypothesis corresponding to w (parameterized by)
- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ a loss function. $\ell(a, b)$

Gradient: $\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(h_w(x_i), y_i)$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial a} \ell(h_w(x_i), y_i) \frac{\partial}{\partial w} h(x_i)$$

Example: LSR (least squares regression)

- $h_w(x_i) = \langle w, x_i \rangle$
- $\ell(a, b) = (a - b)^2$

$$\Rightarrow L(w) = \frac{1}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right)^2$$

Gradient: $\nabla L(w) = \frac{2}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right) x_i$

$$\nabla L(w)_j = \frac{1}{n} \sum_{i=1}^{n} 2 \left(\langle w, x_i \rangle - y_i \right) \frac{\partial}{\partial w_j} \langle w, x_i \rangle$$

$$= \frac{2}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right) x_{ij} \qquad \frac{\partial}{\partial w_j} \sum_{j'=1}^{d} w_{j'} x_{ij'} = x_{ij}$$

Costs $O(n \cdot d)$ time to compute.

(n examples, d dimensions)

$\Rightarrow$ Time complexity of GD $\approx O\left(\frac{n \cdot d}{\varepsilon}\right)$

$n \cdot d$ very bad in many applications.

Stochastic Gradient Descent:

idea: efficiently estimate $\nabla f(w)$.

- Let $G(w)$ be an estimator for $\nabla f(w)$

  i.e. $\mathbb{E}[G(w)] = \nabla f(w)$ (unbiased estimator)

  $\rightarrow$ $G(w)$ a random vector in $\mathbb{R}^d$

e.g. LSR: $\nabla L(w) = \frac{2}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i) x_i$

Could use: • $G(w) = 2(\langle w, x_{i^*} \rangle - y_{i^*}) x_{i^*}$ where i^* sampled u.a.r.

or • for $k \ll n$:

$$G(w) = \frac{2}{k} \sum_{i \in S} (\langle w, x_i \rangle - y_i) x_i \quad \text{where}$$

$S \subseteq [n]$, $|S| = k$ random.

SGD: (using estimator G)

1) Choose w_0
2) for $i = 1, \dots, T$:

   let $w_{i+1} = w_i - \eta G(w_i)$

How "good" is G_t? Small variance is good.

- $\mathrm{Var}(G_t) = \mathbb{E}\left[\|G_t(w) - \mathbb{E}[G_t(w)]\|_2^2\right]$

$$\mathbb{E}\left[\|G_t - \mathbb{E}[G_t]\|_2^2\right]$$

$$= \mathbb{E}\left[\sum_{j=1}^{d}\left(G_{t,j} - \mathbb{E}[G_{t,j}]\right)^2\right] = \sum_{j=1}^{d}\mathbb{E}\left[\left(G_{t,j} - \mathbb{E}[G_{t,j}]\right)^2\right]$$

$$= \sum_{j=1}^{d}\mathrm{Var}(G_{t,j})$$

Thm (SGD convergence):

f convex & β-smooth, $\eta \le 1/\beta$,

$\mathrm{Var}(G_t(w)) \le \sigma^2$

$$\Rightarrow \mathbb{E}\left[f(\bar{w}_k)\right] \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta \cdot k} + \eta\sigma^2$$

(error term introduced by using G_t)

$\bar{w}_k = \frac{1}{k}(w_1 + \cdots + w_k)$

Remarks:

- SGD is what is most commonly used in practice (variants).
- Can get "accelerated" version of SGD (as in NAGD) just by replacing $\nabla f(w_k)$ with $G_t(w_k)$ in NAGD.

Proof of SGD convergence:

useful things: $\forall u, v \in \mathbb{R}^d$:

- convexity: $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle$
- smoothness: $f(u) \leq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\beta}{2} \|u - v\|_2^2$
- cosine formula: $2\langle u, v \rangle = \|u\|_2^2 + \|v\|_2^2 - \|u - v\|_2^2$
- SGD update: $w_{k+1} = w_k - \eta G(w_k)$
- unbiased estimator: $\mathbb{E}[G(w)] = \nabla f(w)$

$$\Rightarrow \quad \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\nabla f(w)\|_2^2 \leq \sigma^2$$

$$\sigma^2 \geq \mathbb{E}\left[\|G(w) - \mathbb{E}[G(w)]\|_2^2\right] = \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\mathbb{E}[G(w)]\|_2^2$$

$$= \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\nabla f(w)\|_2^2$$

Clm 1: (SGD function decrease ineq) $\forall k$ (from smoothness)

$$\star \quad \mathbb{E}\left[f(w_{k+1})\right] \le \mathbb{E}\left[f(w_k)\right] - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\eta\sigma^2}{2}$$

pf:

(smoothness) $f(w_{k+1}) \le f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{\beta}{2}\|w_{k+1} - w_k\|_2^2$

$(w_{k+1} = w_k - \eta G(w_k))$ $= f(w_k) + \langle \nabla f(w_k), -\eta G(w_k) \rangle + \frac{\beta\eta^2}{2}\|G(w)\|_2^2$

$$\sigma^2 \ge \mathbb{E}\left[\|G(w) - \mathbb{E}[G(w)]\|_2^2\right] = \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\mathbb{E}[G(w)]\|_2^2$$
$$= \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\nabla f(w)\|_2^2$$

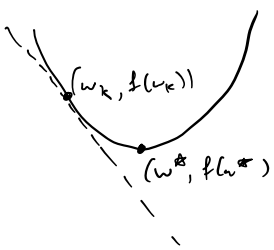
$$\mathbb{E}\left[f(w_{k+1})\right] \le \mathbb{E}\left[f(w_k)\right] - \eta\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\beta\eta^2}{2}\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\beta\eta^2\sigma^2}{2}$$

$(\eta \le 1/\beta,\ \beta \le 1/\eta)$ $\le \mathbb{E}\left[f(w_k)\right] - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\eta\sigma^2}{2}$. □

Now, similar to GD analysis, let's use convexity of f:

$\star\star$ Convexity: $f(w_k) \le f(w^*) + \langle \nabla f(w_k), w_k - w^* \rangle$

$\left(-\langle \nabla f(w_k), w_k - w^* \rangle \le f(w^*) - f(w_k)\right)$



$\star\star\star$ $\frac{1}{2\eta}\mathbb{E}\left[\|w_{k+1} - w^*\|_2^2 - \|w_k - w^*\|_2^2\right]$ (Relates point distance to the gradient)

Twist ineq. $= \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] - \mathbb{E}\left[\langle \nabla f(w_k), w_k - w^* \rangle\right] + \frac{\eta\sigma^2}{2}$

Pf: $\|w_{k+1} - w^*\|_2^2 = \|w_k - w^* - \eta G(w_k)\|_2^2 = \|w_k - w^*\|_2^2 + \eta^2 \|G(w_k)\|_2^2 - 2\eta \langle w_k - w^*, G(w_k)\rangle$

(a = $w_k - w^*$, v = $G(w_k)$)

$$\Rightarrow \mathbb{E}\left[\|w_{k+1} - w^*\|_2^2 - \|w_k - w^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|\nabla f(w_k)\|_2^2 + \eta^2\sigma^2 - 2\eta \mathbb{E}\left[\langle \nabla f(w_k), w_k - w^*\rangle\right]\right]$$

multiply by $1/2\eta$ ... $\square$

main proof continues...

$(-\mathbb{E}[\langle \nabla f(w_k), w_k - w^* \rangle] \le f(w^*) - \mathbb{E}[f(w_k)])$

$\star\star\star + \star\star \Rightarrow$

$$\frac{1}{2\eta}\mathbb{E}\left[\|w_{k+1} - w^*\|_2^2 - \|w_k - w^*\|_2^2\right] \le \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + f(w^*) - \mathbb{E}[f(w_k)] + \frac{\eta\sigma^2}{2}$$

$$\le f(w^*) - \mathbb{E}[f(w_{k+1})] + \eta\sigma^2 \quad (\text{by } \star)$$

$\Rightarrow \forall i = 1, \dots, k$: (Telescoping sum trick)

$$\mathbb{E}[f(w_i)] \le f(w^*) - \frac{1}{2\eta}\mathbb{E}\left[\|w_i - w^*\|_2^2 - \|w_{i-1} - w^*\|_2^2\right] + \eta\sigma^2$$

$$\Rightarrow \sum_{i=1}^{k} \mathbb{E}[f(w_i)] \le k \cdot f(w^*) - \frac{1}{2\eta}\left(\|w_k - w^*\|_2^2 - \|w_0 - w^*\|_2^2\right) + k\eta\sigma^2$$

$$\le k \cdot f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta} + k\eta\sigma^2$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} f(w_i)\right] \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta} + \eta\sigma^2$$

Now, f convex $\Rightarrow f\left(\frac{1}{k}\sum_{i=1}^{k} w_i\right) \le \frac{1}{k}\sum_{i=1}^{k} f(w_i)$

and recall that SGD output is $\bar{w}_k = \frac{1}{k}\sum_{i=1}^{k} w_i$. $\square$

Concluding Remarks:

Thm (SGD convergence):

f convex & β-smooth, $\eta \le 1/\beta$,

$\mathrm{Var}(G(w)) \le \sigma^2$

$$\Rightarrow \mathbb{E}\left[f(\bar{w}_k)\right] \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta \cdot k} + \eta\sigma^2$$

(equality)

η large: good: allows us to converge faster (as long as $\eta \le 1/\beta$)

η small: good: mitigates error of estimator

Choosing step-size:

in standard GD: just want to pick η as large as possible (i.e. $\eta = 1/\beta$)

in SGD: have to balance the 2 error terms

general guideline: choose $\eta \approx 1/\sqrt{k}$

$$\eta = 1/\sqrt{k} \Rightarrow \mathbb{E}\left[f(\bar{w}_k)\right] \le f(w^*) + \frac{1}{\sqrt{k}}\left(\frac{\|w_0 - w^*\|_2^2}{2} + \sigma^2\right)$$

$$\eta = \frac{\|w_0 - w^*\|_2}{\sigma\sqrt{2k}} \Rightarrow \mathbb{E}\left[f(\bar{w}_k)\right] \le f(w^*) + \frac{\sqrt{2}\cdot\sigma\|w_0 - w^*\|_2}{\sqrt{k}}$$

~ in practice, step-size choice can be more nuanced depending on the specific application.

... or use $\eta_t = 1/\sqrt{t}$ (step-size decreases over time)

Constrained Optimization: So far we have only considered optimizing over all of $\mathbb{R}^d$.

i.e. $\boxed{\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; L(w)}$

In many situations we want w to satisfy some constraints.

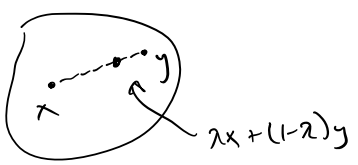
i.e. want w to lie in some bounded region of $\mathbb{R}^d$.

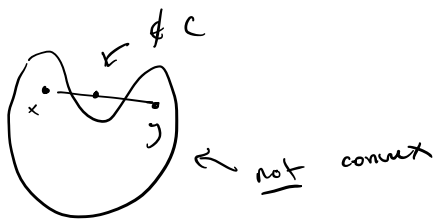
It turns out we can adapt GD / SGD / NAGD to this situation as long as the constrained region is a convex set.

Defn: $C \subseteq \mathbb{R}^d$ a convex set if

$\forall x, y \in C, \; \frac{x+y}{2} \in C$ (midpoint also in C)

equiv: $\forall x, y \in C, \; \forall \lambda \in [0,1], \; \lambda x + (1-\lambda) y \in C$





not convex

Goal: compute $\boxed{\underset{w \in C}{\operatorname{argmin}} \; L(w)}$

Projected Gradient Descent:

Projection: $\operatorname{Proj}_C(y) = \underset{x \in C}{\operatorname{argmin}} \; \|x - y\|_2$ (closest point in C to y)

PGD update: $w_{k+1} = \operatorname{Proj}_C(w_k - \eta(\nabla f(w_k)))$

↑ or $G(w_k)$ in case of SGD

Remark : Everything known for GD/SGD/NAGD extends to PGD as long as C is a convex set.