Stochastic Gradient Descent.

Recap from last time :

- S a convex open set
- $f : S \to \mathbb{R}$ convex fct.
- f ρ-Lipschitz : has subgradients $\|v_t\| \leq \rho$.

GD : $w^1 = 0$

$$w^{t+1} = w^t - \eta \underbrace{v(w^t)}_{\nabla f(w^t) \text{ a subgradient}} \quad , t = 1 \dots T$$

Thm : Let $\bar{w} = \underset{\|w\| \leq B}{\operatorname{argmin}} f(w)$.

For T steps and rate $\eta = \frac{B}{\rho\sqrt{T}}$ we have

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}$$

where $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} f(w^t)$.

Often it is not possible to access the true gradient or subgradient. Moreover some kind of stochasticity in the steps through S might be beneficial. Stochastic gradient descent is a popular method to avoid costly or impossible gradient computation and at the same time it builds in stochasticity.

Definition : Stochastic gradient.

A stochastic gradient of f at point z is a random variable V_z such that

$$\mathbb{E}(V_z) \in \partial f(z).$$

In words the expected value of the stochastic gradient must be a subgradient. If f is differentiable $\mathbb{E}(V_z) = \nabla f(z)$.

Stochastic Gradient Descent Algorithm:

- Let $\eta > 0$ rate and $T \in \mathbb{N}$ number of steps.
- SGD algorithm is the following stochastic process:

a) initialize $w^1 = 0$

b) for $t = 1 \dots T$ do

$$\underline{w}^{t+1} = \underline{w}^t - \eta \underline{v}^t$$

where $\underline{v}^t$ is a random vector s.t

$$\mathbb{E}(\underline{v}^t \mid v^1 v^2 \dots v^{t-1}) \in \partial f(w^t)$$

c) output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^t$.

Note: at each time t we choose a stochastic gradient afresh given past history. This is an unbiased estimator of $\nabla f(w^t)$ (if f is differentiable).

Theorem : approach of SGD to optimal value.

Let $B, \rho > 0$. Let f be convex.

Assume that with probability 1 the stochastic gradients satisfy $\|v^t\| \leq \rho$ (this replaces Lipshitzed).

Let $w^* = \underset{\|w\| \leq B}{\operatorname{argmin}} f(w)$ minimizer in Ball$(0, B)$.

For T steps and $\eta = \frac{B}{\rho\sqrt{T}}$ we have

$$0 \leq \mathbb{E}\left(f(\bar{w})\right) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

where the expectation is over $v^1 v^2 \dots v^T$.

Proof of Thm.

The proof follows the steps of the one for usual GD.

$$\mathbb{E}\left(f(\bar{w})\right) - f(w^*) = \mathbb{E}\left(f(\bar{w}) - f(w^*)\right)$$

$$\leq \quad \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left(f(w^t) - f(w^*)\right)\right)$$

$\leq$ convexity of f.

$$\leq \quad \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\langle v^t, w^t - w^*\rangle\right)$$

$\leq$ to be shown below

$$\leq \frac{BS}{\sqrt{T}}$$

[by some proof then last time because with prob 1 $\|v^t\| \leq S$.]

Now we must show (*) because v^t is not the "true deterministic" subgradient or gradient.

To be shown now:

$$\mathbb{E}_{v_1 \dots v_T}\left[\frac{1}{T}\sum_{t=1}^{T} f(w^t) - f(w^*)\right] \leq \mathbb{E}_{v_1 \dots v_T}\left[\frac{1}{T}\sum_{t=1}^{T}\langle v^t, w^t - w^*\rangle\right]$$

$$\text{Left. hand. side} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v_1 \dots v^T}\left(f(w^t) - f(w^*)\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v^1 \dots v^{t-1}}\left(f(w^t) - f(w^*)\right) \quad (*)$$

$\uparrow$

w^t depends on history $v^1 \dots v^{t-1}$

(since recall $w^t = w^{t-1} - \eta v^{t-1}$, $\dots w^2 = w^1 - \eta v^1$)

Now by convexity of f for fixed $v^1 \dots v^{t-1}$:

$$f(w^*) \geq f(w^t) + \langle \underbrace{\nabla f(w^t)}, w^* - w^t \rangle$$

any vector in $\partial f(w^t)$

$$= \mathbb{E}\left[v^t \mid v^1 \dots v^{t-1}\right]$$

by definition of stochastic v^t gradient

$$\Rightarrow \quad f(w^t) - f(w^*) \leq \left\langle \mathbb{E}\left(v^t \mid v^1 \dots v^{t-1}\right), w^t - w^* \right\rangle$$

$$= \mathop{\mathbb{E}}_{v^t \mid v^1 \dots v^{t-1}} \langle v^t, w^t - w^* \rangle$$

↑

depends only on $v^1 \dots v^{t-1}$.

Replacing in (*) we obtain :

$$\text{Left\_hand\_side} \leq \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_1 \dots v_{t-1}} \left(\mathop{\mathbb{E}}_{v_t \mid v_1 \dots v_{t-1}} \left(\langle v^t, w^t - w^* \rangle \right) \right)$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_1 \dots v_t} \left(\langle v^t, w^t - w^* \rangle \right)$$

$$= \mathop{\mathbb{E}}_{v_1 \dots v_T} \left(\frac{1}{T} \sum_{t=1}^{T} \langle v^t, w^t - w^* \rangle \right)$$

$$= \text{Right\_hand side.}$$

■

Learning with SGD.

We discuss the application of the SGD algorithm to learning. Recall we would like to minimize

$$L_{\mathcal{D}}(\underline{w}) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(\underline{w}, z)]$$

Here we changed notation $h : \mathcal{X} \to \mathcal{Y}$ for $\underline{w}$. Imagine $h(\underline{x}) = y$ represents a NN with weights $\underline{w}$.

Since we do not know $\mathcal{D}$ we cannot optimize $L_{\mathcal{D}}(\underline{w})$ directly and we proposed to replace it by the empirical risk $L_S(\underline{w})$ as a proxy.

Here we propose something different: let v^t be an unbiased estimator of $\nabla_w L_{\mathcal{D}}(w)$

i.e $\mathbb{E}[v^t / v^1 \dots v^{t-1}] = \nabla_w L_{\mathcal{D}}(w^t)$

We can take :

$$\boxed{v^t = \nabla_w \ell(w^t, z)}.$$

Interpretation : at time t consider a fresh sample $z \sim \mathcal{D}$ and compute $\nabla_w \ell(w^t, z)$ for this sample. [This can be seen as replacing $\nabla_w L_S(w) = \frac{1}{N} \sum_{i \in S} \nabla_w \ell(w, z_i)$ by one term taken at random at time t : $\nabla_w \ell(w, z_t)$]

SGD algorithm for learning:

Fix $\eta > 0$ rate, $T \in \mathbb{N}$ number of steps

a) initialize $w^1 = 0$.

b) for $t = 1 \dots T$ do

- sample $z^t \sim \mathcal{D}$ afresh.
- pick $v^t \in \partial \ell(w^t, z^t)$ or $v^t = \nabla_w \ell(w^t, z^t)$
- update $w^{t+1} = w^t - \eta \nabla_w \ell(w^t, z^t)$

c) output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^t$.

Corrolary: Let $\ell(w, z)$ be a convex and ρ-Lipshitz loss fct $\ell(\cdot, z)$.

Then $\forall \varepsilon > 0$ if we run SGD with $T \geq \frac{B^2 \rho^2}{\varepsilon^2}$ and $\eta = \frac{B}{\rho\sqrt{T}}$ then we have:

$$\mathbb{E}\left(L_{\mathcal{D}}(\bar{w})\right) \leq \min_{w \in Ball(0,B)} L_{\mathcal{D}}(w) + \varepsilon.$$

Proof follows from previous (exercise). #.