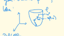
Stochastic Gradient Descent

§1 Classical Gradient Descent

Setup: $f: S \to \mathbb{R}$, differentiable. This means $\nabla f: S \to \mathbb{R}^n$, $S \subseteq \mathbb{R}^n$

Goal: Minimize f i.e. find $w^* \in S$ s.t. $w^* \in \operatorname{argmin} f(w)$.

Idea:



Def (GD):
Input $\eta > 0$, $T \in \mathbb{N}$, $f: S \to \mathbb{R}$
Initialize $w^1 \in S$
For $t = 1 \dots T$
  $w^{t+1} = w^t - \eta \nabla f(w^t)$
Return $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^t$

Other options: Return w^T or best performing w^t s.t. $f(w^t) \le f(w^s)$, $t \le T$.

§2 Subgradient

We want to generalize GD to non-diff. functions.

Recall: If $f: S \to \mathbb{R}$ is convex then for all point $w \in S$ if f is differentiable in w then

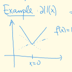
$$\forall u \in S: \quad f(u) \ge \underbrace{f(w) + \langle u - w, \nabla f(w)\rangle}_{\text{tangent hyperplane of } f \text{ at } w}$$

Lem: Let S be convex, open. Then a function $f: S \to \mathbb{R}$ is convex
$\iff \forall w \in S\ \exists v \in \mathbb{R}^n\ \forall u \in S: f(u) \ge f(w) + \langle u - w, v\rangle$.

Def: Let $f: S \to \mathbb{R}$ be a function. A vector $v \in \mathbb{R}^n$ satisfying
$\forall u \in S: f(u) \ge f(w) + \langle u - w, v\rangle$
at a given point $w \in S$ is called subgradient of f at w. The set of all subgradients is denoted with $\partial f(w)$.

Facts: i) f is convex, then $\partial f(w) \ne \emptyset$ for $w \in S$.
ii) f is convex and differentiable then $\partial f(w) = \{\nabla f(w)\}$

Example: $\partial |x| = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ \{-1\} & x < 0 \end{cases}$, $x \in \mathbb{R}$



§3 Lemma 14.1

For later convergence theorems we need

Lem 14.1: Let $v_1, \dots, v_T \in \mathbb{R}^n$, then any algorithm with initialization $w^1 = 0$ and update rule
$$w^{t+1} = w^t - \eta v_t, \quad \eta > 0$$
satisfies for all $w^* \in \mathbb{R}^n$
$$\sum_{t=1}^{T} \langle w^t - w^*, v_t\rangle \le \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
In particular for every $B, \rho > 0$ if $\|v_1\|, \dots, \|v_T\| \le \rho$ and $w^* \in \bar{B}_B(0)$ then with $\eta = \frac{B}{\rho}\frac{1}{\sqrt{T}}$ we have
$$\frac{1}{T}\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle \le \frac{B\rho}{\sqrt{T}}$$

Proof: Using the polarization identities
$$\langle u, v\rangle = \frac{1}{2}\left(\|u\|^2 + \|v\|^2 - \|u - v\|^2\right)$$
$$\Rightarrow \langle w^t - w^*, v_t\rangle = \frac{1}{\eta}\langle w^t - w^*, \eta v_t\rangle$$
$$= \frac{1}{2\eta}\left(-\|w^t - w^* - \eta v_t\|^2 + \|w^t - w^*\|^2 + \eta^2\|v_t\|^2\right)$$
$$= \frac{1}{2\eta}\left(\|w^t - w^*\|^2 - \|w^{t+1} - w^*\|^2\right) + \frac{\eta}{2}\|v_t\|^2$$
Summing over t we get (telescopic sum)
$$\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle = \frac{1}{2\eta}\sum_{t=1}^{T}\left(\|w^t - w^*\|^2 - \|w^{t+1} - w^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$= \frac{1}{2\eta}\left(\|w^1 - w^*\|^2 - \|w^{T+1} - w^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$\le \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

§4 Stochastic Gradient Descent

Def (SGD):
Input $\eta > 0$, $T \in \mathbb{N}$, $f: S \to \mathbb{R}$
Initialize $w^1 = 0$
For $t = 1 \dots T$
  Choose v_t according to some prob. distribution s.t. $\mathbb{E}[v_t | w^t] \in \partial f(w^t)$
  $w^{t+1} = w^t - \eta v_t$
Return $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^t$

Thm 14.8: Let $B, \rho > 0$, $f: S \to \mathbb{R}$ convex ($S \subseteq \mathbb{R}^n$).
Let $w^* \in \operatorname{argmin}_{\|w\| \le B} f(w)$. Assume that SGD runs with T iterations and stepsize $\eta = \frac{B}{\rho}\frac{1}{\sqrt{T}}$ and assume that
$\|v_1\|, \dots, \|v_T\| \le \rho$ a.s.
Then $\mathbb{E}f(\bar{w}) - f(w^*) \le \frac{B\rho}{\sqrt{T}}$. Therefore for given $\varepsilon > 0$ to achieve
$$\mathbb{E}f(\bar{w}) - f(w^*) \le \varepsilon$$
we need to run $T \ge \left(\frac{B\rho}{\varepsilon}\right)^2$ iterations of SGD.

Proof: We use $v_{1:T} = v_1, \dots, v_T$. By Jensen's inequality
$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T}\left(f(w^t) - f(w^*)\right)$$
$$\Rightarrow \mathbb{E}_{v_{1:T}}[f(\bar{w}) - f(w^*)] \le \mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(w^t) - f(w^*)\right)\right]$$
Using Lemma 14.1
$$\mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle\right] \le \frac{B\rho}{\sqrt{T}}$$
This means that it is enough to show
$$\mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(w^t) - f(w^*)\right)\right] \le \mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle\right]$$
Using linearity of expectation
$$\mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{v_{1:T}}[\langle w^t - w^*, v_t\rangle]$$
Next recall law of total expectation:
For α, β random variables, g some function then $\mathbb{E}_\alpha[g(\alpha)] = \mathbb{E}_\beta\mathbb{E}_\alpha[g(\alpha)|\beta]$
Now put $\alpha = v_{1:t}$, $\beta = v_{1:t-1}$. We obtain
$$\mathbb{E}_{v_{1:T}}[\langle w^t - w^*, v_t\rangle] = \mathbb{E}_{v_{1:t}}[\langle w^t - w^*, v_t\rangle]$$
$$= \mathbb{E}_{v_{1:t-1}}\mathbb{E}_{v_{1:t}}[\langle w^t - w^*, v_t\rangle | v_{1:t-1}] = \mathbb{E}_{v_{1:t-1}}[\langle w^t - w^*, \mathbb{E}_{v_t}[v_t | v_{1:t-1}]\rangle]$$
But now w^t is determined: $w^t = -\eta(v_1 + \dots + v_{t-1})$
It follows
$$\mathbb{E}_{v_t}[v_t | v_{1:t-1}] = \mathbb{E}_{v_t}[v_t | w^t] \in \partial f(w^t)$$
$$\Rightarrow \mathbb{E}_{v_{1:t-1}}[\langle w^t - w^*, \mathbb{E}_{v_t}[v_t | v_{1:t-1}]\rangle] \ge \mathbb{E}_{v_{1:t-1}}[f(w^t) - f(w^*)]$$
$$\Rightarrow \mathbb{E}_{v_{1:T}}[\langle w^t - w^*, v_t\rangle] \ge \mathbb{E}_{v_{1:T}}[f(w^t) - f(w^*)]$$
Now summing over $t \le T$ and dividing by T gives the desired inequality.
To achieve
$$\mathbb{E}f(\bar{w}) - f(w^*) \le \frac{B\rho}{\sqrt{T}} \le \varepsilon \iff T \ge \left(\frac{B\rho}{\varepsilon}\right)^2 \quad \square$$

§5 Stochastic Gradient Descent for Risk Minimization

In Learning Theory we want to minimize the function
$$L_D(w) = \mathbb{E}_{z \sim D}[\ell(w, z)]$$
SGD allows us to directly minimize L_D.
For simplicity we assume that $\ell(\cdot, z)$ is differentiable for all $z \in Z$.
We construct the random v_t as follows: Sample $z \sim D$
$$v_t = \nabla \ell(w^t, z)$$
$\Rightarrow$ where the gradient is taken w.r.t. w. Interchanging integration and gradient
$$\mathbb{E}[v_t | w^t] = \mathbb{E}_z[\nabla \ell(w^t, z)]$$
$$= \nabla\mathbb{E}_z[\ell(w^t, z)] = \nabla L_D(w^t) \in \partial L_D(w^t).$$
The same argument can be applied to the subgradient case.
Let $v_t \in \partial \ell(w^t, z)$ for sampled $z \sim D$.
Then by def
$$\forall u: \quad \ell(u, z) - \ell(w^t, z) \ge \langle u - w^t, v_t\rangle$$
$$\Rightarrow \quad L_D(u) - L_D(w^t) \ge \mathbb{E}_z[\langle u - w^t, v_t\rangle | w^t]$$
$$= \langle u - w^t, \mathbb{E}_z[v_t | w^t]\rangle$$
$$\Rightarrow \mathbb{E}[v_t | w^t] \in \partial L_D(w^t).$$

Def (SGD for minimizing L_D):
Input: $\eta > 0$, $T \in \mathbb{N}$
Initialize $w^1 = 0$
For $t = 1 \dots T$
  Sample $z \sim D$
  Pick $v_t \in \partial \ell(w^t, z)$
  $w^{t+1} = w^t - \eta v_t$
Return $\bar{w} = \frac{1}{T}\sum_{t} w^t$

By theorem 14.8

Cor: Let $B, \rho > 0$, L_D convex such that
$$\|\nabla \ell(w^1, z)\|, \dots, \|\nabla \ell(w^T, z)\| \le \rho \quad \text{a.s.}$$
Let $w^* \in \operatorname{argmin}_{\|w\| \le B} f(w)$. Then to achieve
$$\mathbb{E}f(\bar{w}) - f(w^*) \le \varepsilon$$
for given $\varepsilon > 0$ we need to run SGD with stepsize $\eta = \frac{B}{\rho}\frac{1}{\sqrt{T}}$ and $T \ge \left(\frac{B\rho}{\varepsilon}\right)^2$ iterations.