Stochastic Gradient Descent:

Parameters: $\eta > 0$, $T > 0$

Initialization: $w^{(1)} = 0$

Steps: $t = 1, \ldots T$

Choose $v^{(t)}$ from a distribution so that $\mathbb{E}[v^{(t)} | w^{(t)}] \in \partial f(w^{(t)})$.

$$w^{(t+1)} = w^{(t)} - \eta v^{(t)}.$$

Output: $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

Theorem: Let $B, \rho > 0$. Let f be a convex function and let $w^* \in \operatorname{argmin}_{w:\|w\| \le B} f(w)$. Run SGD for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Then

$$\mathbb{E}[f(\bar{w})] - f(w^*) \le \frac{B\rho}{\sqrt{T}}.$$

Note: This is the same result we had before except that now the statement is about $\mathbb{E}[f(\bar{w})]$ instead of $f(\bar{w})$.

Assume that wp 1

$\|V^{(t)}\| \leq \rho$.

Proof: The proof proceeds along the same lines as before.

$$\mathbb{E}[f(\bar{w}) - f(w^*)] \overset{\text{convexity}}{\leq}$$

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

Note that the expectation is over all the choices of $v^{(t)}$, $t = 1, \dots T-1$. We have

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

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

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

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

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

$$\leq \frac{B\rho}{\sqrt{T}}$$ identical proof to before.

Note: There is of course still the variance!

Note: X, Y rvs; $\mathcal{F}$ filtration
Y is $\mathcal{F}$-measurable

$$\mathbb{E}[XY|\mathcal{F}] = Y\,\mathbb{E}[X|\mathcal{F}]$$

Lemma: Let $V = \operatorname{argmin}_{x \in \mathcal{H}} \|x - w\|^2$,

where $\mathcal{H}$ is closed and convex.
Then $\forall u \in \mathcal{H}$

$$\|V - u\|^2 \leq \|w - u\|^2.$$

See exercise.

The previous proof works more or less as is because of the

Adding a projection step.

We require w^* to be in a ball of radius B. But we cannot guarantee that $\bar{w}$ fulfills this.

We can add a projection step.

$$w^{(t+\frac{1}{2})} = w^{(t)} - \eta \nabla f(w^{(t)})$$

$$w^{(t+1)} = \underset{w \in \|\cdot\| \leq B}{\operatorname{argmin}} \|w - w^{(t+\frac{1}{2})}\|$$

We get the same convergence guarantees in this way. But now $\bar{w}$ is also in the same ball.

Variable step size.

$$\eta_t = \frac{B}{\rho\sqrt{t}}$$

After averaging.

Output average over last αT steps.

Strongly convex functions.

$$f(v) \geq f(w) + \langle v-w, \nabla f(w)\rangle + \frac{\varepsilon}{2}\|w-v\|^2.$$

$$f(\alpha w + \bar{\alpha} v) \leq \alpha f(w) + \bar{\alpha} f(v) - \frac{\varepsilon}{2}\alpha\bar{\alpha}\|w-v\|^2.$$

Learning with SGD

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

Note:

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

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

Therefore, as long as we pick a new sample for each step of the SGD algorithm we do get an unbiased gradient!

One-hidden layer NNs and SGD: (67)

$$\hat{f}(x;\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i)$$

N number of neurons in hidden layer.

$x \in \mathbb{R}^d$: d neurons in input layer

$$\sigma_*(x;\theta_i) = a_i\,\sigma(\langle w_i, x\rangle + b_i)$$

$\sigma: \mathbb{R} \to \mathbb{R}$; e.g.: $\sigma(x) = \underbrace{\frac{1}{1+e^{-x}}}$ sigmoid

derivative 1 at $x=0$

$\sigma(x) = \max\{0, x\}$ ReLU