Online and Stochastic Gradient Descent, Multi-Armed Bandits

Dimitris Fotakis

School of Electrical and Computer Engineering
National Technical University of Athens, Greece

Online Gradient Descent

Online Gradient Descent

Input: convex set S, $\boldsymbol{w}_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play $\boldsymbol{w}_t$, get cost function $f_t : S \to \mathbb{R}$, incur cost $f_t(\boldsymbol{w}_t)$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Online Gradient Descent

Online Gradient Descent

Input: convex set S, $\boldsymbol{w}_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play $\boldsymbol{w}_t$, get cost function $f_t : S \to \mathbb{R}$, incur cost $f_t(\boldsymbol{w}_t)$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Ignore project step for the analysis: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
Also we let $\boldsymbol{v}_t = \nabla f_t(\boldsymbol{w}_t)$ for brevity.

Online Gradient Descent

Online Gradient Descent

Input: convex set S, $\boldsymbol{w}_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play $\boldsymbol{w}_t$, get cost function $f_t : S \to \mathbb{R}$, incur cost $f_t(\boldsymbol{w}_t)$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Ignore project step for the analysis: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
Also we let $\boldsymbol{v}_t = \nabla f_t(\boldsymbol{w}_t)$ for brevity.

Using convexity $f_t(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_t) + \nabla f_t(\boldsymbol{w}_t)(\boldsymbol{w}^* - \boldsymbol{w}_t)$, we get that:

$$\sum_{t=1}^{T} (f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq \sum_{t=1}^{T} \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) \tag{1}$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow \\ \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) &= \frac{1}{2\eta}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2) + \frac{\eta}{2}\|\boldsymbol{v}_t\|^2 \qquad (2)\end{aligned}$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow\end{aligned}$$

$$\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) = \tfrac{1}{2\eta}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \qquad (2)$$

Substituting (2) in (1) results in a telescopic sum. So, we get that:

$$\begin{aligned}\text{Regret}_{OGD} = \sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) &\leq \frac{\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \frac{\eta T G^2}{2} \overset{\eta = \frac{B}{G\sqrt{T}}}{=} BG\sqrt{T}\end{aligned}$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow\end{aligned}$$

$$\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) = \tfrac{1}{2\eta}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \qquad (2)$$

Substituting (2) in (1) results in a telescopic sum. So, we get that:

$$\begin{aligned}\text{Regret}_{OGD} = \sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) &\leq \frac{\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \frac{\eta T G^2}{2} \overset{\eta = \frac{B}{G\sqrt{T}}}{=} BG\sqrt{T}\end{aligned}$$

Similar regret with step $\eta_t = \frac{B}{G\sqrt{t}}$, because $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$.

α-Strongly Convex Functions

Using α-strong convexity

$$f_t(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_t) + \nabla f_t(\boldsymbol{w}_t)(\boldsymbol{w}^* - \boldsymbol{w}_t) + \frac{\alpha}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_t\|^2 ,$$

we get that:

$$2\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq \sum_{t=1}^{T}\left(2\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) - \alpha\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\right) \quad (3)$$

α-Strongly Convex Functions

Using α-strong convexity

$$f_t(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_t) + \nabla f_t(\boldsymbol{w}_t)(\boldsymbol{w}^* - \boldsymbol{w}_t) + \frac{\alpha}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_t\|^2,$$

we get that:

$$2\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq \sum_{t=1}^{T}\left(2\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) - \alpha\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\right) \quad (3)$$

Substituting (2) in (3), we get that:

$$\begin{aligned}
2\text{Regret}_{OGD} &= 2\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \\
&\leq \sum_{t=1}^{T}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha\right) + \sum_{t=1}^{T}\eta_t\|\boldsymbol{v}_t\|^2 \\
&\overset{\eta_t = 1/(\alpha t)}{\leq} 0 + \frac{G^2(1+\ln T)}{\alpha}
\end{aligned}$$

Learning Using Stochastic Gradient Descent

- In learning problems, we want to solve $\min_{\boldsymbol{w} \in H} L_{\mathcal{D}}(\boldsymbol{w})$, where $L_{\mathcal{D}}(H) = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)]$ and $\ell(\boldsymbol{w}, z)$ is the loss on z under $\boldsymbol{w} \in H$.
- Instead of Empirical Risk Minimization $\min_{\boldsymbol{w} \in H} L_S(\boldsymbol{w})$, we try to minimize $L_{\mathcal{D}}(\boldsymbol{w})$ directly using SGD and the fact that:

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

Learning Using Stochastic Gradient Descent

- In learning problems, we want to solve $\min_{\boldsymbol{w} \in H} L_{\mathcal{D}}(\boldsymbol{w})$, where $L_{\mathcal{D}}(H) = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)]$ and $\ell(\boldsymbol{w}, z)$ is the loss on z under $\boldsymbol{w} \in H$.
- Instead of Empirical Risk Minimization $\min_{\boldsymbol{w} \in H} L_S(\boldsymbol{w})$, we try to minimize $L_{\mathcal{D}}(\boldsymbol{w})$ directly using SGD and the fact that:

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

- We cannot calculate $\nabla L_{\mathcal{D}}(\boldsymbol{w})$, because we do not know $\mathcal{D}$.
- But expected value of $\nabla \ell(\boldsymbol{w}, z)$, over $z \sim \mathcal{D}$, is $\nabla L_{\mathcal{D}}(\boldsymbol{w})$
- We use $\nabla \ell(\boldsymbol{w}, z)$, for sample $z \sim \mathcal{D}$, which is an **unbiased estimator** of the gradient.

Stochastic Gradient Descent

Stochastic Gradient Descent

Input: convex set H, $\boldsymbol{w}_1 \in H$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Random sample $z_t \sim \mathcal{D}$, $\boldsymbol{v}_t = \nabla \ell(\boldsymbol{w}_t, z_t)$ and $f_t(\boldsymbol{w}_t) = \boldsymbol{v}_t \boldsymbol{w}_t$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{v}_t$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in H} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Return $\bar{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$

Stochastic Gradient Descent

Stochastic Gradient Descent

Input: convex set H, $\boldsymbol{w}_1 \in H$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Random sample $z_t \sim \mathcal{D}$, $\boldsymbol{v}_t = \nabla \ell(\boldsymbol{w}_t, z_t)$ and $f_t(\boldsymbol{w}_t) = \boldsymbol{v}_t \boldsymbol{w}_t$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{v}_t$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in H} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Return $\bar{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$

Analysis based on OGD, with $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$.

- SGD tries to minimize $f(\boldsymbol{w}) = L_{\mathcal{D}}(\boldsymbol{w})$
- Note that $\nabla f(\boldsymbol{w}_t) = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\nabla \ell(\boldsymbol{w}_t, z)] = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\boldsymbol{v}_t] = \nabla f_t(\boldsymbol{w}_t)$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(w) = v_t w$

- SGD tries to minimize $f(w) = L_{\mathcal{D}}(w)$
- Note that $\nabla f(w_t) = \mathbb{E}\text{xp}_{z\sim\mathcal{D}}[\nabla \ell(w_t, z)] = \mathbb{E}\text{xp}_{z\sim\mathcal{D}}[v_t] = \nabla f_t(w_t)$

$$\mathbb{E}\text{xp}_{z_1,\ldots,z_T}[f(\bar{w})] - f(w^*) \leq \mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \frac{f(w_t) - f(w^*)}{T}\right]$$

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

$$= \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} v_t(w_t - w^*)\right]$$

$$= \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} f_t(w_t) - f_t(w^*)\right]$$

$$\leq \frac{\text{Regret}_{OGD}(T)}{T} \leq \frac{BG}{\sqrt{T}}$$