First Order Methods for Online Convex Optimization

Tao Li

OptML Seminar, ISE, Lehigh Universty

October 11, 2019

1. Recap on OCO
2. Online Gradient Descent
3. Example: Stochastic Gradient Descent
4. Online Gradient Descent for Strongly Convex Functions

Recap on OCO

Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of the correct answers to previous questions and possibly additional available information.

- Goal: minimize the cumulative loss suffered along its run
- Process: deduce information from previous rounds to improve its predictions on present and future questions

Recap on OCO

The difference between the real cumulative loss and this minimum cumulative loss in hindsight is defined as regret:

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)$$

- If regret grows linearly, the player is not learning.
- If regret grows sub-linearly, $R(T) = o(T)$, the player is learning and its prediction accuracy is improving.

$$\frac{1}{T}\left(\sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x_t)\right) \to 0, \quad T \to \infty$$

Online Gradient Descent

Algorithm 1 Online Gradient Descent (OGD) Algorithm

Input: convex set $\mathcal{K}$, $T, \mathrm{x}_1 \in \mathcal{K}$, step sizes $\{\eta_t\}$
for $k = 1, \ldots, T$ **do**
  Play x_t and observe cost $f_t(\mathrm{x}_t)$
  Update and project:

$$\mathrm{x}_{t+1} = \prod_{\mathcal{K}} \left(\mathrm{x}_t - \eta_t \nabla f_t(\mathrm{x}_t)\right)$$

end for

Online Gradient Descent

- In each iteration, the algorithm takes a step from the previous point in the direction of the gradient of the previous cost. This step may result in a point outside of the underlying convex set. In such cases, the algorithm projects the point back to the convex set.
- The regret attained by the algorithm is sub-linear.

Online Gradient Descent

Theorem

Algorithm 1 with step size $\eta_t = \frac{D}{G\sqrt{t}}$ *guarantees*

$$regret_T = \sum_{t=1}^{T} f_t(\mathrm{x}_t) - \min_{\mathrm{x}^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathrm{x}^\star) \leq \frac{3}{2} GD\sqrt{T}$$

where

$$D : D = \max_{x,y \in \mathcal{K}} \|x - y\|, \text{diameter of } \mathcal{K}$$

$$G : \|\nabla f_t\| \leq G, \text{bound on gradient norm}$$

Online Gradient Descent

Let $\mathrm{x}^\star \in \arg\min_{\mathrm{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathrm{x})$. Define $\nabla_t \triangleq \nabla f_t\left(\mathrm{x}_t\right)$.
By convexity

$$f_t\left(\mathrm{x}_t\right) - f_t\left(\mathrm{x}^\star\right) \leq \nabla_t^\top \left(\mathrm{x}_t - \mathrm{x}^\star\right) \tag{1}$$

By the Pythagorean theorem:

$$\left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2 = \left\|\prod_{\mathcal{K}} \left(\mathrm{x}_t - \eta_t \nabla_t\right) - \mathrm{x}^\star\right\|^2 \leq \left\|\mathrm{x}_t - \eta_t \nabla_t - \mathrm{x}^\star\right\|^2 \tag{2}$$

$$\left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2 \leq \left\|\mathrm{x}_t - \mathrm{x}^\star\right\|^2 + \eta_t^2 \left\|\nabla_t\right\|^2 - 2\eta_t \nabla_t^\top \left(\mathrm{x}_t - \mathrm{x}^\star\right) \tag{3}$$

$$2\nabla_t^\top \left(\mathrm{x}_t - \mathrm{x}^\star\right) \leq \frac{\left\|\mathrm{x}_t - \mathrm{x}^\star\right\|^2 - \left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2}{\eta_t} + \eta_t G^2 \tag{4}$$

Online Gradient Descent

Plug (4) into (1) we have,

$$
\begin{aligned}
2\left(\sum_{t=1}^{T} f_t(\mathrm{x}_t) - f_t(\mathrm{x}^*)\right) &\leq 2\sum_{t=1}^{T} \nabla_t^\top (\mathrm{x}_t - \mathrm{x}^*) \\
&\leq \sum_{t=1}^{T} \frac{\|\mathrm{x}_t - \mathrm{x}^*\|^2 - \|\mathrm{x}_{t+1} - \mathrm{x}^*\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq \sum_{t=1}^{T} \|\mathrm{x}_t - \mathrm{x}^*\|^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \leq 3DG\sqrt{T}
\end{aligned}
$$

Stochastic Optimization

Stochastic problem

$$\min_{x \in \mathcal{K}} f(x)$$

- f is a convex function, $\mathcal{K}$ is a convex domain.
- Access to a noisy gradient $\tilde{\nabla}_t$

$$\mathbb{E}[\tilde{\nabla}_t] = \nabla f(x_t), \mathbb{E}[\|\tilde{\nabla}_t\|^2] \leq G^2. \quad (5)$$

- Define linear loss function $f_t(x) = \tilde{\nabla}_t^\top x$. Applying OGD to f_t, obtain SGD algorithm.
- From regret bound of OGD to convergence rates of SGD.

Stochastic Optimization

Algorithm 2 Stochastic Gradient Descent

Input: f, $\mathcal{K}$, T, $x_1 \in \mathcal{K}$, step size $\{\eta_t\}$
for $k = 1, \ldots, T$ **do**
    Generate $\tilde{\nabla}_t$ s.t. (5)
    Update and project

$$x_{t+1} = \prod_{\mathcal{K}} (x_t - \eta_t \tilde{\nabla}_t)$$

end for
return $\tilde{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

Regret Bound to Convergence Rate

Theorem

Algorithm 2 with step size $\eta = \frac{D}{G\sqrt{T}}$ *has*

$$\mathbb{E}[f(\tilde{x}_T)] \leq f(x^*) + \frac{3GD}{2\sqrt{T}}$$

Regret Bound to Convergence Rate

Proof:

$$
\begin{aligned}
\mathbb{E}[f(\tilde{x}_T)] - f(x^*) \leq & \mathbb{E}[\frac{1}{T}\sum_t f(x_t)] - f(x^*) \\
\leq & \frac{1}{T}\mathbb{E}[\sum_t \nabla f(x_t)^\top (x_t - x^*)] \\
= & \frac{1}{T}\mathbb{E}[\sum_t \tilde{\nabla}_t^\top (x_t - x^*)] \\
\leq & \frac{\text{regret}_T}{T} \\
\leq & \frac{3GD}{2\sqrt{T}}
\end{aligned}
$$

Online Gradient Descent for Strongly Convex Functions

Theorem

For α-strongly convex loss functions, Algorithm 1 with step sizes $\eta_t = \frac{1}{\alpha t}$ has

$$regret_T \leq \frac{G^2}{2\alpha}(1 + \log T)$$

Online Gradient Descent for Strongly Convex Functions

Applying the definition of α-strong convexity to the pair of points $\mathrm{x}_t, \mathrm{x}^*$, we have

$$2\left(f_t\left(\mathrm{x}_t\right) - f_t\left(\mathrm{x}^\star\right)\right) \leq 2\nabla_t^\top\left(\mathrm{x}_t - \mathrm{x}^\star\right) - \alpha\left\|\mathrm{x}^\star - \mathrm{x}_t\right\|^2 \tag{6}$$

By the Pythagorean theorem:

$$\left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2 = \left\|\prod_{\mathcal{K}}\left(\mathrm{x}_t - \eta_t\nabla_t\right) - \mathrm{x}^\star\right\|^2 \leq \left\|\mathrm{x}_t - \eta_t\nabla_t - \mathrm{x}^\star\right\|^2$$

$$\begin{aligned} \left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2 &\leq \left\|\mathrm{x}_t - \mathrm{x}^\star\right\|^2 + \eta_t^2\left\|\nabla_t\right\|^2 - 2\eta_t\nabla_t^\top\left(\mathrm{x}_t - \mathrm{x}^\star\right) \\ 2\nabla_t^\top\left(\mathrm{x}_t - \mathrm{x}^\star\right) &\leq \frac{\left\|\mathrm{x}_t - \mathrm{x}^\star\right\|^2 - \left\|\mathrm{x}_{t+1} - \mathrm{x}^\star\right\|^2}{\eta_t} + \eta_t G^2 \end{aligned} \tag{7}$$

Online Gradient Descent for Strongly Convex Functions

Plug (7) into (6) we have,

$$
\begin{aligned}
&2\sum_{t=1}^{T}\left(f_t\left(\mathrm{x}_t\right)-f_t\left(\mathrm{x}^{\star}\right)\right) \\
&\leq \sum_{t=1}^{T}\left\|\mathrm{x}_t-\mathrm{x}^{\star}\right\|^2\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\alpha\right)+G^2\sum_{t=1}^{T}\eta_t \\
&=0+G^2\sum_{t=1}^{T}\frac{1}{\alpha t} \\
&\leq \frac{G^2}{\alpha}(1+\log T)
\end{aligned}
$$