Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Victor Ye Dong
Email: yd2470@columbia.edu

SCRIBE

9

Projected Gradient Descent-Continued

1. Projected Gradient Descent

1.1 Recap for algorithm

Algorithm 1: Projected Gradient Descent

for $t = 0, 1, 2...$ **do**
  $y_{t+1} = x_t - \eta \nabla f(x_t)$
  $x_{t+1} = \Pi_x(y_{t+1})$
end
return some combination of $x_0, ..., x_T$

1.2 Example of projections

If $K = \{x \in \mathbb{R} | \; \|x\|_2 \leq R\}$, then,

$$\Pi_K(y) = \begin{cases} y, \text{if } ||y||_2 \leq R \\ \frac{Ry}{||y||_2}, \text{if } ||y||_2 \leq R \end{cases} \tag{1..1}$$

If $K = \{x \in \mathbb{R} | \; \|x\|_\infty \leq R\}$, then,

$$\Pi(y) = \begin{cases} y, \text{if } ||y||_\infty \leq R \\ R, \text{if } ||y||_\infty > R \\ -R, \text{if } ||y||_\infty < -R \end{cases} \tag{1..2}$$

1.3 Project Gradient Descent for α-strongly convex and L-Lipschitz f

We will use a version of projected gradient descent with different step sizes used in different iterations. In iteration t, we use step size η_t. I.e., the update in iteration t is $x_{t+1} = \Pi_K(x_t - \eta_t \nabla_t f(x_t))$. Defining $y_{t+1} = x_t - \eta_t \nabla_t f(x_t)$, we have, as in previous lectures:

$$||x_{t+1} - x^*||^2 \leq ||y_{t+1} - x^*||^2 = ||x_t - x^*||^2 + \eta_t^2 ||\nabla f(x_t)||^2 - 2\eta_t \nabla f(x_t)(x_t - x^*) \tag{1..3}$$

$\Rightarrow$

$$\begin{aligned} f(x_t) - f(x^*) &\leq \frac{1}{2\eta_t}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) + \frac{\eta_t}{2}||\nabla f(x_t)||^2 - \frac{\alpha}{2}||x_t - x^*||^2 \\ &\leq \frac{1}{2\eta_t}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) + \frac{\eta_t}{2}L^2 - \frac{\alpha}{2}||x_t - x^*||^2 \\ &\leq \frac{1}{2}(\frac{1}{\eta_t} - \alpha)(||x_t - x^*||^2) - \frac{1}{2\eta_t}||x_{t+1} - x^*||^2 + \frac{\eta_t}{2}L^2 \end{aligned} \tag{1..4}$$

Notice that here we cannot follow the previous ways to remove the term of x_t, x_{t+1} by simple summing up LHS and RHS. However we can still make the RHS telescope by setting $\frac{1}{\eta_t} - \alpha = \frac{1}{\eta_{t+1}}$. One choice of η_t which ensures this is $\eta_t = \frac{1}{\alpha(t+1)}$. Then

$$\begin{aligned} \sum_{t=0}^{T-1} f(x_t) - f(x^*) = &\sum_{t=1}^{T-1}(\frac{1}{2}(\frac{1}{\eta_t} - \alpha) - \frac{1}{2\eta_{t-1}})||x_t - x^*||^2 + \\ &\frac{1}{2}(\frac{1}{\eta_0} - \alpha)||x_0 - x^*||^2 - \frac{1}{2\eta_{T-1}}||x_T - x^*||^2 + \sum_{t=0}^{T}\frac{\eta_t}{2}L^2 \end{aligned} \tag{1..5}$$

Notice that $\frac{1}{\eta_0} - \alpha = 0$. Dropping the non-positive term $-\frac{1}{2\eta_{T-1}}||x_T - x^*||^2$ and using Jensen's inequality, we have

$$\begin{aligned} f(\tfrac{1}{T}\textstyle\sum_{t=0}^{T-1} x_t) - f(x^*) &\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2\alpha(t+1)}L^2 \\ &\leq \frac{L^2}{2\alpha} \cdot \frac{\ln(T) + 1}{T}. \end{aligned} \tag{1..6}$$