Online Machine Learning Homework

Assignment 3

Student: Lei Feng, `please-help@each.other`
Student ID: 123456
Lecturer:

Problem 1: Strong Convexity

Strongly convex functions have a number of important properties that make them particularly nice to use as regularizers. We investigate some of them here. For all of these, assume the functions are from $\mathbb{R}^n \to \mathbb{R}$, and strong convexity is with respect to an arbitrary norm $\|\cdot\|$. Unless otherwise specified, assume the strong convexity holds on some convex set $\mathcal{W}$.

(a) Suppose $f : \mathbb{R}^n \to \mathbb{R}$, is σ-strongly convex on a convex set $\mathcal{W}$ (possibly $\mathbb{R}^n$). Show that f is strongly convex on any convex $\mathcal{W}' \subseteq \mathcal{W}$.

(b) Let f be σ-strongly convex, and let h be α-strongly-convex. Show that $c(x) = f(x) + h(x)$ is $\sigma + \alpha$-strongly-convex. An important corollary is that if f is σ-strong-convex and h is an arbitrary convex function, then their sum is also σ strongly-convex.

(c) Suppose f is 1-strongly-convex. Show that $h(x) = \alpha f(x)$ is α-strongly-convex for $\alpha \in [0, \infty)$

(d) Let f be σ-strongly-convex on a convex set $\mathcal{W}$. Show that f has a unique minimizer $w^* \in \mathcal{W}$.

(a) *Proof.* Since $\mathcal{W}' \subseteq \mathcal{W}$, $\forall\, x, x' \in \mathcal{W}'$ and $\forall\, \theta \in [0, 1]$, we have $x, x' \in \mathcal{W}$.

Additionally f is σ-strongly convex on a convex set $\mathcal{W}$, we then have

$$f(\theta x + (1-\theta)x') \le \theta f(x) + (1-\theta) f(x') - \frac{\sigma}{2}\theta(1-\theta)\|x - x'\|^2$$

Hence f is also σ-strongly convex on $\mathcal{W}'$. □

(b) *Proof.* By the definition of strong convexity, for f we have

$$f(\theta x + (1-\theta)x') \le \theta f(x) + (1-\theta) f(x') - \frac{\sigma}{2}\theta(1-\theta)\|x - x'\|^2 \quad (1)$$

For h we have

$$h(\theta x + (1-\theta)x') \le \theta h(x) + (1-\theta) f(x') - \frac{\alpha}{2}\theta(1-\theta)\|x - x'\|^2 \quad (2)$$

Adding Eq.1 and Eq.2, we then have

$$c(x) = c(\theta x + (1-\theta)x') \le \theta c(x) + (1-\theta) f(x') - (\frac{\sigma}{2} + \frac{\alpha}{2})\theta(1-\theta)\|x - x'\|^2$$

Thus, $c(x)$ is $\sigma + \alpha$-strongly-convex. □

(c) *Proof.* Since $f(x)$ is 1-strongly-convex. We have

$$f(\theta x + (1-\theta)x') \le \theta f(x) + (1-\theta) f(x') - \frac{1}{2}\theta(1-\theta)\|x - x'\|^2$$

Then, for $h(x) = \alpha f(x)$, we have

$$h(\theta x + (1-\theta)x') = \alpha f(\theta x + (1-\theta)x') \le \alpha\theta f(x) + \alpha(1-\theta) f(x') - \frac{\alpha}{2}\theta(1-\theta)\|x - x'\|^2$$

which can be simplified to

$$h(\theta x + (1-\theta)x') \le \theta h(x) + (1-\theta) h(x') - \frac{\alpha}{2}\theta(1-\theta)\|x - x'\|^2$$

Hence, $h(x) = \alpha f(x)$ is α-strongly-convex for $\alpha \in [0, \infty)$. □

(d) Strong convexity implies that the second derivative is positive definite, thus ensuring a global minimum as the function curves upwards, hence the uniqueness of w^*.

Problem 2: Online Gradient Descent with Strongly Convex Loss Functions

Recall the analysis of the Online Gradient Descent algorithm (see notes for lecture 5)

(a) Prove that if the loss functions f_t are all σ-strongly convex then regret is upper bounded by

$$\frac{1}{2}\|w^*\|^2(\frac{1}{\eta_t} - \sigma) + \frac{1}{2}\sum_{t=2}^{T}\|w_t - w^*\|^2(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_{t+1}\ . \qquad (3)$$

(b) Set $\eta_t = \frac{1}{\sigma t}$ and conclude that

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \le \frac{G^2}{H}(1 + \log T)\ .$$

Note that we obtain a *logarithmic* bound on regret, which is much smaller than a *square-root* bound

(a) *Proof.* Since f_t are all σ-strongly convex, we have

$$f_t(w_t) - f_t(w^*) \ge \langle \nabla f_t(w^*), w_t - w^* \rangle + \frac{\sigma}{2}\|w_t - w^*\|^2.$$

Additionally, w_t updates as

$$w_{t+1} = w_t - \eta_t \nabla f_t(w_t)$$

Then, we have

$$\langle \nabla f_t(w^*), w_t - w^* \rangle \le \frac{1}{2\eta_t}\|w_t - w^*\|^2 + \frac{\eta_t}{2}\|\nabla f_t(w_t)\|^2.$$

and,

$$\|w_{t+1} - w^*\|^2 = \|w_t - \eta_t \nabla f_t(w_t) - w^*\|^2$$

Finally, we have

$$R(T) = \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \le \frac{1}{2}\|w^*\|^2(\frac{1}{\eta_t} - \sigma) + \frac{1}{2}\sum_{t=2}^{T}\|w_t - w^*\|^2(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_{t+1} \ .$$

□

(b) *Proof.* By setting $\eta_t = \frac{1}{\sigma t}$, we have $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \sigma$, then

$$R(T) = \frac{G^2}{2}\sum_{t=1}^{T}\eta_{t+1} = \frac{G^2}{2\sigma}\sum_{t=1}^{T}\frac{1}{t}.$$

Essentially, we are setting the first two terms in Eq.3 to 0.

Since that

$$\sum_{t=1}^{T}\frac{1}{t} \le 1 + \log T$$

thus,

$$R(T) \le \frac{G^2}{2\sigma}(1 + \log T)$$

which follows the form of $\frac{G^2}{H}(1 + \log T)$. □

Problem 3: Implementing FTRL with Proximal and L1 Regularization

We consider the FTRL algorithm with adaptive proximal regularization and an L_1 penalty to introduce sparsity. We consider the unconstrained problem, so the update is

$$w_{t+1} = \arg\min_{w \in \mathbb{R}^n} g_{1:t} \cdot w + t\lambda\|w\|_1 + \sum_{s=1}^{t}\frac{\sigma_s}{2}\|w - w_s\|_2^2$$

Here, the σ_s in $\mathbb{R}^+$ give the strength of each incremental regularization function, and $\lambda \ge 0$ gives the strength of the per-round L_1 penalty.

(a) Consider the $1D$ optimization problem

$$w^* = \arg\min_{w \in \mathbb{R}^n} \frac{a}{2}w^2 + bw + c\|w\|_1 \ ,$$

where $a, b, c \in \mathbb{R}$ are constants and $a, c \ge 0$. Derive a closed-form solution for w^*. Hint: Consider the subdifferential of this objective, and recall that if $0 \in \partial f(w^*)$ then w^* is a minimizer. Your closed-form solution may still contain cases.

(b) Suppose that the σ_t are chosen only as a function of t, for example so $\sigma_{1:t} = \sqrt{t}$, corresponding to a learning rate of $\frac{1}{\sqrt{t}}$. Write pseudocode for the algorithm, using an implementation that only requires storing a single vector in $\mathbb{R}^n$. For simplicity, structure your code like this:

```
/*TODO: Define variables for the state of the algorithm*/
for round t = 1, 2, ... do
    Observe gradient g_t
    /*TODO: Implement the update*/
    /*TODO: Compute and output w_{t+1} */
end for
```

(a) The subdifferential of $|w|$ is:

$$\partial|w| = \begin{cases} \{1\}, & \text{if } w > 0, \\ \{-1\}, & \text{if } w < 0, \\ [-1, 1], & \text{if } w = 0. \end{cases}$$

The subdifferential of the objective function is:

$$\partial\Big(\frac{a}{2}w^2 + bw + c|w|\Big) = aw + b + c \cdot \partial|w|$$

Consider w^* under three conditions:

– If $w^* > 0$: then $\partial|w^*| = 1$. The equation becomes:

$$0 = aw^* + b + c.$$

Solving for w^*:

$$w^* = -\frac{b+c}{a} \tag{4}$$

– If $w^* < 0$: then $\partial|w^*| = -1$. The equation becomes:

$$0 = aw^* + b - c.$$

Solving for w^*:

$$w^* = -\frac{b-c}{a} \tag{5}$$

– If $w^* = 0$: then $\partial|w^*| = [-1, 1]$.

$$0 \in b + c \cdot [-1, 1].$$

This holds if and only if $|b| \leq c$.
Combining them altogether, we have

$$w^* = \begin{cases} -\frac{b+c}{a}, & \text{if } b + c < 0 \\ -\frac{b-c}{a}, & \text{if } b - c > 0 \\ 0, & \text{if } |b| \leq c \end{cases}$$

(b) **Initialize:** $g_{\text{sum}} \leftarrow 0$, $w \leftarrow 0$

for round $t = 1, 2, \ldots$ **do**

Observe gradient g_t

Update: $g_{\text{sum}} = g_{\text{sum}} + g_t$

Compute w_{t+1}:

$$w_{t+1}[i] = \text{sign}(-g_{\text{sum}}[i]) \cdot \max(0, |g_{\text{sum}}[i]| - t\lambda)/(t\sigma_t)$$

Output w_{t+1}

end for