CONVEXITY AND NO REGRET NOTES

LUCAS TUCKER

Abstract. Below are some brief notes/exercises on online learning and convex optimization, drawing from Vishnoi's *Algorithms for Convex Optimization and Orabona's A Modern Introduction to Online Learning.*

Contents

1. Notes

Definition 1.1. We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. We say f is "strictly convex" if the inequality is strict for $x \neq y$.

Lemma 1.2. *For $f : \mathbb{R}^n \to \mathbb{R}^m$ a differentiable function at x_0 defined on $A \subset \mathbb{R}^n$ and $g : D \to \mathbb{R}^k$ differentiable at $f(x_0)$ with range$(f) \subset D$, for $F := g(f(x))$, we have*

$$F'(x_0) = g'(f(x)) \cdot f'(x)$$

Proof. Let $A = f'(x_0)$ and $B = g'(f(x_0))$. Further, for $h \in \mathbb{R}^n, \delta \in \mathbb{R}^m$ given by $\delta = f(x_0 + h) - f(x_0)$, let

$$a(h) = f(x_0 + h) - f(x_0) - Ah$$
$$b(\delta) = g(f(x_0) + \delta) - g(f(x_0)) - B\delta$$

We then have $a(h) =: |h|s(h)$ and $b(\delta) =: |\delta|t(\delta)$ for $|s(h)| = o(|h|)$ and $|t(\delta)| = o(|\delta|)$. Hence,

$$|\delta| = |f(x_0 + h) - f(x_0)| = |Ah + a(h)| \leq |h|(||A|| + |s(h)|)$$

so that

$$\begin{aligned} |F(x_0 + h) - F(x_0) - BAh| &= |b(\delta) + B\delta - BAh| \\ &= |t(\delta)|\delta| + B(\delta - Ah)| = |t(\delta)|\delta| + Ba(h)| \\ &\leq |\delta||t(\delta)| + ||B|||h||s(h)| \end{aligned}$$

We then have

$$\frac{|F(x_0 + h) - F(x_0) - BAh|}{|h|} \leq \frac{|\delta||t(\delta)| + ||B|||h||s(h)|}{|h|}$$

so substituting the inequality for $|\delta|$ yields the following as $|h| \to 0$:

$$\leq (||A|| + |s(h)|)|t(\delta)| + ||B|||s(h)| \to 0$$

$\square$

Lemma 1.3. *For $f : \mathbb{R}^d \to \mathbb{R}$ continuously differentiable and $g : [0,1] \to \mathbb{R}$ defined as*

$$g(t) := f(x + t(y - x))$$

we have

$$\dot{g}(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$$
$$f(y) - f(x) = \int_0^1 \dot{g}(t)dt$$
$$\ddot{g}(t) = (y - x)^T \nabla^2 f(x + t(y - x))(y - x)$$

Proof. Apply the Fundamental Theorem of Calculus with Lemma 1.2 $\square$

Lemma 1.4. *For $f : D \to \mathbb{R}$ a continuously differentiable function over a convex set $D \subset \mathbb{R}^d$, f is convex if and only if, for all $x, y \in D$ we have*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$$

Theorem 1.5. *For $D \subset \mathbb{R}^d$ a convex open domain and $f : D \to \mathbb{R}$ a smooth function, f is convex if and only if*

$$\nabla^2 f(x) \geq 0$$

for all $x \in K$.

Proof. Fix $x \in D$. Since D is open, for any $y \in \mathbb{R}^d$ there exists $t > 0$ such that $x + ty \in D$. Without loss of generality let $||y|| = 1$. Then,

$$0 \leq \frac{1}{t^2}\langle \nabla f(x + ty) - \nabla f(x), ty \rangle = \frac{\langle \nabla f(x + ty), y \rangle - \langle \nabla f(x), y \rangle}{t}$$

so that for $g(t) := f(x + t(y' - x))$ where $y' = y + x$, by Lemma 1.3

$$= \frac{1}{t}(\dot{g}(t) - \dot{g}(0)) = \frac{1}{t}\int_0^t \ddot{g}(\xi)d\xi = \frac{1}{t}\int_0^t \langle \nabla^2 f(x + ty)y, y \rangle$$

hence $H(f)$ is positive semi-definite (y was chosen arbitrarily from the unit ball in $\mathbb{R}^d$).

Conversely, if $H(f) = \nabla^2 f$ is positive semi-definite, then for $h(t) = f(x + t(y - x))$ we have

$$f(y) = h(1) = h(0) + \int_0^1 \dot{h}(t)dt$$
$$= h(0) + \dot{h}(0) + \int_0^1 \dot{h}(t) - \dot{h}(0)dt$$
$$\Rightarrow h(1) - h(0) - \dot{h}(0) = \int_0^1 \int_0^t \ddot{h}(\lambda)d\lambda dt \geq 0$$

so that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq 0$$

hence f is convex. □

Definition 1.6. For fixed $\epsilon > 0$ and norm $||.||$, a differentiable function $f : D \to \mathbb{R}$ where $D \subset \mathbb{R}^d$ is convex, is considered "ϵ-strongly convex with respect to $||.||$" if

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{\epsilon}{2}||y - x||^2$$

Definition 1.7. The Bregman divergence of a function $f : D \to \mathbb{R}$ at $x, y \in D$ is given by

$$D_f(x, y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

Definition 1.8. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \infty$, we define its conjugate $f^* : \mathbb{R}^d \to \mathbb{R} \cup \infty$ as

$$f^*(y) := \sup_{x \in \mathbb{R}^d} \langle y, x \rangle - f(x)$$

for $y \in \mathbb{R}^n$

Definition 1.9. The Online learning setting works as follows: At the t-th round of T many, the algorithm receives an instance $x_t \in \mathcal{X}$ and makes a prediction $\widehat{y_t} \in \mathcal{Y}$. The algorithm then receives the true label $y_t \in \mathcal{Y}$ and calculates a loss $L(\widehat{y_t}, y_t)$ with $L : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$ a loss function. The algorithm seeks to minimize the cumulative loss $\sum_{t=1}^{T} L(\widehat{y_t}, y_t)$ over the T rounds.

2. Exercises

The following exercises are numbered according to Vishnoi's book

3.10 We wish to show that a convex function $f : D \to \mathbb{R}$ is continuous. The definition of convexity gives $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for $x, y \in D$. Suppose without loss of generality that $f(y + \lambda(x - y)) - f(y) \geq 0$ (otherwise $f(y + \lambda(x - y)) - f(y) \geq 0$

$$f(y + \lambda(x - y)) - f(y) \leq \lambda f(x) - \lambda f(y)$$

where y may be chosen in any direction. Then, for $\epsilon > 0$ we choose δ

3.18 Consider the generalized negative entropy function

$$f(x) = \sum_{i=1}^{n} x_i \log(x_i)$$

for $x \in \mathbb{R}^n_{>0}$

a) We wish to find the gradient and Hessian of f. In this case,

$$\nabla f = (\partial_i f) \in \mathbb{R}^{1 \times n},\ i \in [n]$$

where

$$\partial_i f = \frac{\partial f}{\partial x_i} = \log(x_i) + 1$$

and $H(f) \in \mathbb{R}^{n \times n}$ satisfies

$$H(f)_{ij} = \partial_{ij} f = \frac{\partial f}{\partial x_i \partial x_j} = \begin{cases} 0 & i \neq j \\ \frac{1}{x_i} & i = j \end{cases}$$

b) We now wish to show that f is strictly convex. We find that

$$f(\lambda x + (1-\lambda)y) = \sum_{i=1}^{n} (y_i + \lambda(x_i - y_i)) \log(y_i + \lambda(x_i - y_i))$$

while

$$\lambda f(x) + (1-\lambda) f(y) = \sum_{i=1}^{n} \lambda x_i \log(x_i) + (1-\lambda) y_i \log(y_i)$$

hence it suffices to show that $g(t) = t \log(t)$ is strictly convex for $t \in \mathbb{R}$. In particular, we find that $\ddot{g}(t) = \frac{1}{t} > 0$ for $t > 0$, so since each $x_i > 0$ ($x \in \mathbb{R}^n_{>0}$), strict convexity of f follows.

c) However, f is not strongly convex with respect to the ℓ_2 norm, i.e. for fixed $\epsilon > 0$ we have $f(y) - f(x) < \langle \nabla f(x), y - x \rangle + \frac{\epsilon}{2} ||y - x||_2^2$ for some $x, y \in \mathbb{R}^n$. Pick $N \in \mathbb{N}$ large enough such that $\epsilon > \frac{1}{N}$. Note that

$$\langle \nabla f(x), y - x \rangle + \frac{\epsilon}{2} ||y - x||_2^2 = \langle (\log(x_i), ..., \log(x_n)), (y_i - x_i, ..., y_n - x_n)^T \rangle + \frac{\epsilon}{2} \sum_{i=1}^{n} (y_i - x_i)^2$$

$$= \sum_{i=1}^{n} (y_i - x_i)(\log(x_i) + 1) + \frac{\epsilon}{2} \sum_{i=1}^{n} (y_i - x_i)^2$$

Note that, for $t = 1 + \sqrt{\frac{8N}{\epsilon}}$,

$$\frac{\epsilon}{2}(1-t)^2 + \log(t) + 1 - t = 4N + \log\left(1 + \sqrt{\frac{8N}{\epsilon}}\right) - \sqrt{\frac{8N}{\epsilon}}$$

$$\geq 4N - \sqrt{8N^2} = (4 - \sqrt{8})N > 0$$

Then, for $x = \left(1 + \sqrt{\frac{8N}{\epsilon}}, ..., 1 + \sqrt{\frac{8N}{\epsilon}}\right)$, we have

$$0 < \sum_{i=1}^{n} (\log(x_i) + 1 - x_i) + \frac{\epsilon}{2} \sum_{i=1}^{n} (1 - x_i)^2$$

$$\iff -\sum_{i=1}^{n} x_i \log(x_i) < \sum_{i=1}^{n} (1 - x_i)(\log(x_i) + 1) + \frac{\epsilon}{2} \sum_{i=1}^{n} (1 - x_i)^2$$

which, for $y = (1, ..., 1)^T$ is equivalent to

$$f(y) - f(x) < \sum_{i=1}^{n} (y_i - x_i)(\log(x_i) + 1) + \frac{\epsilon}{2} \sum_{i=1}^{n} (y_i - x_i)^2$$

so that the choice of $x = \left(1 + \sqrt{\frac{8N}{\epsilon}}, ..., 1 + \sqrt{\frac{8N}{\epsilon}}\right)^T$ and $y = (1, ..., 1)^T$ suffices.

d) The Bregman Divergence of f in this case is

$$D_f(x, y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$
$$= \sum_{i=1}^{n} (y_i \log(y_i) - x_i \log(x_i)) - \sum_{i=1}^{n} (\log(x_i) + 1)(y_i - x_i)$$
$$= \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} (y_i \log(x_i) + y_i - x_i)$$

However,

$$D_f(y, x) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$
$$= \sum_{i=1}^{n} (x_i \log(x_i) - y_i \log(y_i)) - \sum_{i=1}^{n} (\log(y_i) + 1)(x_i - y_i)$$
$$= \sum_{i=1}^{n} x_i \log(x_i) - \sum_{i=1}^{n} (x_i \log(y_i) + x_i - y_i)$$

so that $D_f(x, y) \neq D_f(y, x)$ for all $x, y \in \mathbb{R}^n_{>0}$.

3. Bibliography

References

[1] http://www.ams.org/publications/authors/tex/amslatex
[2] Francesco Orabona. A Modern Introduction to Online Learning
[3] Nisheeth Vishnoi. Algorithms for Convex Optimization