CS 7545: Machine Learning Theory **Fall 2018**

Lecture 3: Convex Analysis + Deviation Bounds

Lecturer: Jacob Abernethy *Scribes: Samarth Mishra, Rithvik Pudota*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

3.1 Convex Analysis

3.1.1 Review

Definition 3.1 (Bregman Divergence) *Let f be differentiable function, the Bregman Divergence D_f is given by*

$$D_f(\vec{x}, \vec{y}) := f(\vec{x}) - f(\vec{y}) - \langle \nabla f(\vec{y}), \vec{x} - \vec{y} \rangle$$

*f is m-**strongly convex** with respect to $\|\cdot\|$ if for all $\vec{x}, \vec{y} \in \mathrm{dom}(f)$,*

$$D_f(\vec{x}, \vec{y}) \geq \frac{m}{2}\|\vec{x} - \vec{y}\|^2$$

*f is L-**strongly smooth** with respect to $\|\cdot\|$ if for all $\vec{x}, \vec{y} \in \mathrm{dom}(f)$,*

$$D_f(\vec{x}, \vec{y}) \leq \frac{L}{2}\|\vec{x} - \vec{y}\|^2$$

Professor notes the Bregman Divergence behaves like a 'distance', even though it is **not** a metric function.

3.1.2 Fenchel Conjugates

Definition 3.2 (Fenchel Conjugate) *Let f be convex function, the Fenchel conjugate f^* is given as*

$$f^*(\vec{\theta}) := \sup_{\vec{x} \in dom(f)} \langle \vec{x}, \vec{\theta} \rangle - f(\vec{x})$$

where f^* is a convex function

(Exercise)

1. Let $f(\vec{x}) = \frac{1}{2}\|\vec{x}\|_2^2$

$$f^*(\vec{\theta}) = \frac{1}{2}\|\vec{\theta}\|_2^2$$

2. Let $f(\vec{x}) = \frac{1}{2}\vec{x} \cdot (M\vec{x})$, where M is positive semi-definite

$$f^*(\vec{\theta}) = \frac{1}{2}\vec{\theta} \cdot (M^{-1}\vec{\theta})$$

Proof:

$$\begin{aligned} f^*(\vec{\theta}) &= \sup_{\vec{x} \in dom(f)} \langle \vec{x}, \vec{\theta} \rangle - f(\vec{x}) \\ &= \sup_{\vec{x} \in dom(f)} \vec{x} \cdot \vec{\theta} - \frac{1}{2}\vec{x} \cdot (M\vec{x}) \end{aligned}$$

Set $g(\vec{x}) = \vec{x} \cdot \vec{\theta} - \frac{1}{2}\vec{x} \cdot (M\vec{x})$

$$\nabla g(\vec{x}) = \vec{\theta} - M\vec{x}$$

Since $\nabla g(M^{-1}\vec{\theta}) = \vec{0}$, g achieves maximum at $\vec{x} = M^{-1}\vec{\theta}$. As such,

$$\begin{aligned} f^*(\vec{\theta}) &= (M^{-1}\vec{\theta}) \cdot \vec{\theta} - \frac{1}{2}(M^{-1}\vec{\theta}) \cdot (MM^{-1}\vec{\theta}) \\ &= (M^{-1}\vec{\theta}) \cdot \vec{\theta} - \frac{1}{2}(M^{-1}\vec{\theta}) \cdot (\vec{\theta}) \\ &= \frac{1}{2}\vec{\theta} \cdot (M^{-1}\vec{\theta}) \end{aligned}$$

■

3. Let $f(\vec{x}) = \frac{1}{p}\|\vec{x}\|_p^p$, $p > 1$

$$f^*(\vec{\theta}) = \frac{1}{q}\|\vec{\theta}\|_q^q$$

where $\frac{1}{p} + \frac{1}{q} = 1$

Fenchel Conjugate Properties

1. if f is differentible function that is strictly convex and smooth, then for all $\vec{x} \in dom(f)$, $\vec{\theta} \in dom(f^*)$

$$\nabla f(\nabla f^*(\vec{\theta})) = \vec{\theta} \qquad \nabla f^*(\nabla f(\vec{x})) = \vec{x}$$

2. if f is closed, convex function

$$(f^*)^* = f$$

3. if f is differentiable function, then for all $\vec{x}$, $\vec{y} \in dom(f)$,

$$D_f(\vec{x}, \vec{y}) = D_{f^*}(\nabla f(\vec{y}), \nabla f(\vec{x}))$$

4. f is 1-strongly convex w.r.t. $\|\cdot\|$ if and only if f^* is 1-strongly smooth w.r.t. the dual norm $\|\cdot\|_*$

3.1.3 Fenchel-Young Inequality

Theorem 3.3 *Let f be function with fenchel conjugate f^*, for all $\vec{x} \in dom(f)$, $\vec{\theta} \in dom(f^*)$*

$$f(\vec{x}) + f^*(\vec{\theta}) \geq \langle \vec{x}, \vec{\theta} \rangle$$

Proof: Fix $\vec{x} \in dom(f)$, $\vec{\theta} \in dom(f^*)$. Clearly,

$$\sup_{\vec{y} \in dom(f)} \langle \vec{y}, \vec{\theta} \rangle - f(\vec{y}) \geq \langle \vec{x}, \vec{\theta} \rangle - f(\vec{x})$$

$$f^*(\vec{\theta}) \geq \langle \vec{x}, \vec{\theta} \rangle - f(\vec{x})$$
$$f(\vec{x}) + f^*(\vec{\theta}) \geq \langle \vec{x}, \vec{\theta} \rangle$$

■

By the Fenchel-Young Inequality, we can attain Young's Inequality given as

$$\frac{1}{p}\|\vec{x}\|_p^p + \frac{1}{q}\|\vec{\theta}\|_q^q \geq \langle \vec{x}, \vec{\theta} \rangle$$

where $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

3.2 Deviation Bounds

Definition 3.4 (Random Variable) *A random variable is a function $X : \Omega \to \mathbb{R}$ where Ω is a measurable space and the mapping to $\mathbb{R}$ is a probability.*

Some needed concepts for this section :

- Cumulative Distribution Function(CDF) of a random variable X :

$$F(t) = Pr(X \leq t)$$

- Assuming F is differentiable, the PDF of X is

$$f(t) = F'(t)$$

  Note that

$$Pr(a \leq X \leq b) = \int_a^b f(t)dt$$

- Random Variables X and Y are independent if $\forall A, B \subseteq \mathbb{R}$,

$$Pr(X \in A \text{ and } Y \in B) = Pr(X \in A)Pr(Y \in B)$$

- The expectation of X is defined as

$$\mathbb{E}[X] = \int X d\mu$$

  and the variance is defined as

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- **Fact :** If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

(Exercise) Prove that if X and Y are independent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

3.2.1 Markov Inequality

Let X be a random variable, such that $X \geq 0$, then

$$\forall t \quad Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Proof: Define

$$Z_t = \mathbb{1}[X > t] \times t$$

where $\mathbb{1}[]$ is the indicator function [1]

Notice that.

$$\begin{aligned} &\forall t,\ Z_t \leq X \\ \implies &\mathbb{E}[X] \geq \mathbb{E}[Z_t] = t \times \mathbb{E}[\mathbb{1}[X > t]] = t \times Pr(X \geq t) \\ \implies &Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \end{aligned}$$

■

[1] Defining some new notation used: $\mathbb{1}[\text{statement}] = \begin{cases} 1 & \text{if statement true;} \\ 0 & \text{if statement false} \end{cases}$.

3.2.2 Chebyshev's Inequality

Let X be any random variable with bounded mean and variance. Then,

$$Pr(|X - \mu| > t\sigma) \leq \frac{1}{t^2}$$

where mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[(X - \mu)^2]$

Proof: Let $Z = (X - \mu)^2$,

$$\begin{aligned} Pr(|X - \mu| > t\sigma) &= Pr(Z > t^2\sigma^2) && \text{[Since both sides of the inequality are positive, they can be squared]} \\ &\leq \frac{\mathbb{E}[Z]}{t^2\sigma^2} && \text{[Using Markov's Inequality]} \\ &= \frac{1}{t^2} && \text{[Since } \mathbb{E}[Z] = \sigma^2\text{]} \end{aligned}$$

■