CS 7545: Machine Learning Theory **Fall 2019**

Lecture 3: Convex Analysis and Deviation Bounds

Lecturer: Jacob Abernethy *Scribes: Nathaniel Todd, Pol Llado*

3.1 Bregman Divergence Review

Definition 3.1 (Bregman Divergence) *Given a convex, differentiable function $f : \mathbb{U} \to \mathbb{R}$ the* ***Bregman Divergence is defined as***

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

Example: If f is the discrete entropy function, the Bregman divergence is equivalent to the KL Divergence:

$$D_{entropy} := \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \text{ [KL Divergence]}$$

3.1.1 Facts:

1. Bregman Divergence is always positive: $D_f(\vec{x}, \vec{y}) \geq 0$
2. If f is strictly convex, then $D_f(\vec{x}, \vec{y}) = 0$ if and only if $\vec{x} = \vec{y}$
3. f **is μ-strictly convex** with respect to a norm $\|\cdot\|$ if and only if
$$D_f(\vec{x}, \vec{y}) \geq \frac{\mu}{2} \|\vec{x} - \vec{y}\|^2$$
4. f **is β-smooth** with respect to a norm $\|\cdot\|$ if and only if
$$D_f(\vec{x}, \vec{y}) \leq \frac{\beta}{2} \|\vec{x} - \vec{y}\|^2$$

3.1.2 Trivial Fact: Pinsker's Inequality

Pinsker's Inequality is a useful relationship for regularization studies later in the course:

$$KL(p, q) \geq \frac{1}{2} \|p - q\|_1^2 \quad \text{[Pinsker's Inequality]}$$

Proof:

- KL Divergence is 1-Strongly Conxex with respect to the L1 Norm ($\|\cdot\|_1$)
- Bregman Divergence fact 3 above:
$$D_f(\vec{x}, \vec{y}) \geq \frac{\mu}{2} \|\vec{x} - \vec{y}\|^2$$

KL Divergence is a form of Bregman divergence, so if 1-strongly convex then pinsker's Inequality holds:

$$KL(p, q) = D_{entropy}(\vec{x}, \vec{y}) \geq \frac{\mu}{2} \|\vec{x} - \vec{y}\|^2$$

■

This fact is important, because good regularizers are known to have strong convexity for a given norm. Being 1-strongly convex one reason that the L1 Norm is a widely used regularizer in machine learning.

3.2 Fenchel Conjugate

Definition 3.2 (Fenchel Conjugate) *Let f be a convex, twice-differentiable function.* ***The Fenchel conjugate*** *of f is*

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

Claim 3.3 *$f^*(\vec{\theta})$ is also convex*

Proof:

First, let us define an intermediate function:

$$G_x(\vec{\theta}) = \langle \vec{x}, \vec{\theta} \rangle - f(\vec{x})$$

G_x is linear in $\vec{\theta}$, therefore it is convex by definition.

$$f^*(\vec{\theta}) = \sup_{\vec{x} \in dom(f)} G_x(\vec{\theta})$$

we know that the supremum of convex functions is convex, therefore the Fenchel conjugate is convex.

■

3.2.1 Fenchel examples

3.2.1.1 Example 1: 2-Norm

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

3.2.1.2 Example 2: Matrix

Define f as follows where M is a positive definite matrix

$$f(\vec{x}) = \frac{1}{2}\vec{x}^\top M \vec{x}$$

$$f^*(\vec{\theta}) = \sup_{\vec{x}} \langle \vec{x}, \vec{\theta} \rangle - \frac{1}{2}\vec{x}^\top M \vec{x}$$

In order to find the Supremum, we can define $G_\theta(x)$ and find the location at which its gradient is zero.

$$G_\theta(\vec{x}) = \langle \vec{x}, \vec{\theta} \rangle - \frac{1}{2}\vec{x}^\top M \vec{x}$$

Remember from previous mathematics courses that the following is true:

$$C(\vec{x}) = \frac{1}{2}\vec{x}^\top M \vec{x}$$

$$\nabla C(\vec{x}) = M\vec{x}$$

We can use that fact to find the gradient of $G_{\vec{\theta}}(\vec{x})$ as follows:

$$G_\theta(\vec{x}) = \langle \vec{x}, \vec{\theta} \rangle - \frac{1}{2}\vec{x}^\top M \vec{x}$$

$$\nabla_x G_\theta(\vec{x}) = \vec{\theta} - M\vec{x} = 0 \quad \textit{[Solving for Supremum]}$$

$$\vec{x} = M^{-1}\vec{\theta}$$

Now that we have solved for this value of x, we can plug back into the supremum equation. Remember that we can define the inner product as $\langle \vec{x}, \vec{y} \rangle = \vec{y}^\top \vec{x}$

$$\text{Combine: } f^*(\vec{\theta}) = \sup_{\vec{x}} \langle \vec{x}, \vec{\theta} \rangle - \frac{1}{2}\vec{x}^\top M \vec{x}\ ,\ \vec{x} = M^{-1}\vec{\theta} \quad \text{[At Supremum]}$$

$$f^*(\vec{\theta}) = \langle M^{-1}\vec{\theta}, \vec{\theta} \rangle - \frac{1}{2}(M^{-1}\vec{\theta})^\top M (M^{-1}\vec{\theta})$$

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

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

3.2.1.3 p-norm

$$f(\vec{x}) = \frac{1}{p}\|\vec{x}\|_p^p \quad \text{then} \quad f^*(\vec{\theta}) = \frac{1}{q}\|\vec{\theta}\|_q^q$$

$$\text{for } \frac{1}{p} + \frac{1}{q} = 1 \quad \text{and } p > 1$$

Leave proving this as excercise for practice, it is easy to see how Example 1 is a subcase of this exampe.

3.2.2 Fenchel Conjugate Facts

1. *If f is a closed and convex function, then:*

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

2. *If f is strictly convex and differentiable for all x in the domain of f and all* θ *in the domain of* f^*

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

3. *Let f be differentiable and strictly convex, then:*

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

4. ***f is*** μ***-strongly convex*** *with respect to a given norm* $\|\cdot\|$ ***if and only if*** f^* *is* $\frac{1}{\mu}$*-smooth with respect to its dual norm* $\|\cdot\|_*$

3.3 Fenchel-Young Inequality

Claim 3.4 *For a given $\vec{x} \in$ dom(f), $\vec{\theta} \in$ dom(f*) it follows that:*

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

Proof:

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

Because we are taking a supremum over y, we know that any given x plugged in will be less than or equal to the supremum, so we can write:

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

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

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

■

Corollary 3.5 *Combining Claim 3.4 and the Fenchel Fact 3.2.1.3, we obtain the following:*

$$f(\vec{x}) = \frac{1}{p}\|\vec{x}\|_p^p \quad \text{then} \quad f^*(\vec{\theta}) = \frac{1}{q}\|\vec{\theta}\|_q^q$$

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

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

$$\text{for } \frac{1}{p} + \frac{1}{q} = 1 \quad \text{and } p > 1$$

3.4 Deviation Bounds

3.4.1 Random Variable Review

- *A random variable, X, is a measurable function from a σ algebra, Ω, to the set of real numbers, $\mathbb{R}$ where Ω is a sample space and the mapping to $\mathbb{R}$ is a probability.*
- *The expectation of a random variable X, $E[X]$, is defined as*

  $$\int X(\Omega) d\mu$$

  where μ is the underlying measurement.
- *The variance, Var(X), is defined as*

  $$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

- *If X and Y are independent then $E[XY] = E[X]E[Y]$*

- *Distribution Functions of a Random Variable X:*
  The ***Cumulative Distribution Function****, $F(t)$, is defined as*

$$Pr(X \leq t)$$

  The ***Probability Density Function****, $f(t)$ is defined as*

$$F'(t)^{\ddagger}$$

- *The probability that X is between a and b, $Pr(a \leq X \leq b)$ is the area under the PDF:*

$$\int_a^b f(t)dt$$

(Exercise) *Prove that if X and Y are independent then it follows that:*

$$Var(X+Y) = Var(X) + Var(Y))$$

3.4.2 Markov's Inequality

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

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

Proof: *Let*

$$Z_t = {}^{\dagger}1[X > t]t$$

For all t:

$$Z_t \leq X$$
$$E[X] \geq E[Z_t]$$
$$E[Z_t] = tE[1[X > t]] = tPr(X \geq t)$$
$$E[X] \geq tPr(X \geq t)$$
$$Pr(X \geq t) \leq \frac{E[X]}{t}$$

■

3.4.3 Chebyshev's Inequality

Let X be a random variable with bounded mean, $E[X] = \mu$, and bounded variance, σ^2 :
In class the professor presented this version of Chebyshev's Inequality:

$$Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}$$

‡Assuming $F(t)$ is differentiable

†1[input] is the indicator function which outputs 1 if the input is true and 0 if input if false

Proof:

$$Pr[|X-\mu| \geq t] = Pr[|X-\mu|^2 > t^2]$$

Using Markov's Inequality:

$$Pr[|X-\mu| \geq t] \leq \frac{E[(X-\mu)^2]}{t^2}$$

$$E[(X-\mu)^2] = \sigma^2$$

$$Pr[|X-\mu| \geq t] \leq \frac{\sigma^2}{t^2}$$

■

However, the following version can be found in the book and has a nearly identical proof:

$$Pr[|X-\mu| \geq t'\sigma] \leq \frac{1}{t'^2}$$

Additionally, by letting $t = t'\sigma$ *we can see that the former inequality is trivially equivalent to this latter version.*