WHAT’S A LEGENDRE-FENCHEL TRANSFORMATION AND WHY SHOULD I CARE?

R. HASEGAWA AND B. KARMAKAR

Abstract. Several frequently used inequalities - Cauchy-Schwarz, Hölder’s and Minkowski’s inequalities - are often “pulled out of thin air” in mathematics and probability. We will see that these inequalities can be derived from the general structure of Legendre-Fenchel transformations and the associated notion of convex duality by way of an intermediate integral inequality known as Young’s Inequality. What is also interesting is that, in many situations, the structure of convex duality gives us a powerful and direct path to controlling tail probabilities without the need to invoke the arguments of Markov, Cauchy-Schwarz or Minkowski’s inequality. We briefly discuss this tool, which is known as the Cramér-Chernoff method.

1. Legendre-Fenchel Transformation

In the first part of this section, in which we follow the approach by Panagiotopoulos [1985], we will establish some technical notations necessary to discuss the details of Legendre-Fenchel transformation theory in its most general form. In the latter part of the section, we aim to specialize these notions to the real line in order to provide some intuition through simple geometry and plots.

Let \( X \) be a Banach space (which is a normed linear space over a field \( \mathbb{C} \) or \( \mathbb{R} \)) with an associated norm \( \| \cdot \| \). The collection of all continuous linear functional \((X \to \mathbb{R})\) on \( X \) is called the dual space of \( X \) and is denoted by \( X^* \). For \( x^* \in X^* \) and \( x \in X \), we denote the evaluation of the function \( x^*(x) \) by the notation \( \langle x^*, x \rangle \). The pairing \( \langle \cdot, \cdot \rangle \) is known as the duality pairing in the literature.

Let us fix a function \( f : X \to \mathbb{R} \) and look at the affine spaces \( \langle x^*, \cdot \rangle - c \) where \( x^* \in X^* \) and \( c \in \mathbb{R} \). To have a picture in mind, we can think of this affine space as a hyperplane passing through a certain point at the centre \( (x = 0) \) with slope along \( x^* \). Then we are looking for an affine space that ‘supports’ the function \( f \). Technically, for fixed \( x^* \in X^* \) we are asking for best \( c \) such that,

\[
  f(x) \geq \langle x^*, \cdot \rangle - c. \ \forall x \in X.
\]

This leads us to define a conjugate function of \( f \) as,

\[
  f^*(x^*) = \sup_{x \in X} \{ \langle x^*, x \rangle - f(x) \}.
\]
This is called the Legendre-Fenchel transformation\(^1\) of the function. We will now state a few results without proofs but in later part when \(X = \mathbb{R}\) we will see simple arguments for the results.

**Lemma 1.** For any function \(f\) its conjugate \(f^c : X^* \to \mathbb{R}\) is a convex function.

Let us denote by \(\mathcal{C}(X)\) the collection of all real valued functions defined on \(X\) which are pointwise supremum of affine spaces \(\langle x^*, \cdot \rangle - c, \ c \in \mathbb{R}\). We have encountered this collection in the proof of Jensen’s inequality. The following lemma should come to mind immediately.

**Lemma 2.** The set of functions \(\mathcal{C}(X)\) is exactly the set of all real valued convex and lower semi-continuous functions defined on \(X\).

The detailed proof consists of careful analysis and is avoided. Now given a function \(g : X \to \mathbb{R}\), we can look for a function \(f \in \mathcal{C}(X)\) ‘close’ to \(g\). Let’s consider \(f\) as the pointwise supremum of all affine spaces less than \(g\). Then \(f\) is called the convex envelope of the function \(g\). The name convex envelope is explained by the fact that \(f\) is convex and certainly \(f(x) \leq g(x)\) for all \(x \in X\). Using the geometry of the class \(\mathcal{C}\) from Lemma 1 we arrive at the following theorem.

**Lemma 3.**

\[
f^{cc} := (f^c)^c = f \iff f \in \mathcal{C}(X).
\]

After understanding the definition of the convex conjugation our inequality of interest follows immediately from the Equation 1; for any \(f : X \to \mathbb{R}\)

\[
f(x) + f^c(x^*) \geq \langle x^*, x \rangle, \ \forall x \in X, \ \forall x^* \in X^*.
\]

Moreover equality above holds for the pair \(x\) and \(x^*\) if and only if

\[
f(x_1) - f(x) \geq \langle x^*, x_1 - x \rangle \ \forall x_1 \in X.
\]

Which is known as saying \(x^*\) is a sub gradient of \(f\) at the point \(x\). We shall come back to this inequality in Section 2.

1.1. **Legendre-Fenchel transformation on the real line.** When we consider the special case of \(X = \mathbb{R}\), an immediate simplification is that we \(X^* = \mathbb{R}\) (where equality is in the sense of the isometric map \(k \mapsto k \times \cdot \in \mathbb{R}^\ast\)) and the duality pairing \(\langle \cdot, \cdot \rangle\) is simply the product of two numbers. Therefore the definition of the Legendre-Fenchel transformation in (1) of a function \(f : \mathbb{R} \to \mathbb{R}\) provides us with another real function function on the real line,

\[
f^c(k) = \sup_{x \in \mathbb{R}} \{kx - f(x)\}.
\]

Figure 1 delivers some insight into how we can find the convex conjugate of a function \(f\). At this point let us formally write down the definition of sub gradient of a function on real line,

**Definition 4.** We say that a function \(f : \mathbb{R} \to \mathbb{R}\) has a supporting line at \(x\) if there is a \(k \in \mathbb{R}\) such that,

\[
f(x_1) \geq f(x) + k(x_1 - x), \ \forall x_1 \in \mathbb{R}.
\]

All \(a \in \mathbb{R}\) satisfying the above inequality is called sub gradient of the function \(f\) at the point \(x\).

\(^1\)In literature we find various different names identifying this transformation, e.g Legendre transformation, Fenchel-Young transformation, Convex conjugate transformation.
The graph of the convex conjugate $f^c$ is in some sense dual to that of $f$. This duality is made precise by the following lemma.

**Lemma 5.** If $f$ admits a supporting line at $x$ with subgradient $k$, then $f^c$ admits a supporting line with subgradient $x$.

A pictorial ‘proof’ of this lemma can be seen in Figure 2, illustrating the dual relationship between the shapes of $f$ and $f^c$. In words we can say that when $f$ is steeper then $f^c$ is flatter, when $f$ has sharp angle (i.e. is not differentiable) then $f^c$ has a linear section, and so on.

We use the same figure to illustrate another fact. Since $f$ is convex in Figure 2 we know from Lemma 3 that $f^{cc} = f$. But if we consider a function $g$ which is same as $f$ everywhere except on a interval on the right of $x$ as shown by the dotted line then the function is no longer convex. In this case, by Lemma 1 we still have $g^c = f^c$ and then $g \geq f^{cc} = f$, thus $f$ is the convex envelop of $g$.

## 2. Inequalities

The following inequality was first introduced by Young [1912].

**Theorem 6** (Young’s Inequality). Let $f : [0, \infty) \rightarrow [0, \infty)$ be strictly increasing, continuous and satisfy $f(0) = 0$. Then for $a, b \in [0, \infty)$

$$ab \leq \int_0^a f(t) \, dt + \int_0^b f^{-1}(t) \, dt$$

with equality holding iff $b = f(a)$.

Traditionally, proofs of Theorem 6, such as the one in Tolsted [1964], proceed by appealing to the reader’s geometric intuition that the epigraph and hypo-graph of a function constrained to the box $[0, a] \times [0, b]$ add up to $ab$. This proof method is intuitively satisfying. Alternatively, from no more than an observation about the definition of convex conjugates (see Inequality (2)) one lands upon a succinct analytic proof of Young’s Inequality. This is by far not the only analytic proof of

---

**Figure 1.** Graphical interpretation of $f^c$ from the plot of $f$. The graph of the convex conjugate $f^c$ is in some sense dual to that of $f$. This duality is made precise by the following lemma.
Young’s Inequality and the proof given in Diaz and Metcalf [1970] is a nice example of a fully rigorous analytic proof.

Proof. Let \( g(x) = \int_0^x f(t) \, dt \) and recall from (2) that \( ab \leq g(a) + g'(b) \). The proof simply requires that we evaluate the expression for \( g'(b) \). Since \( f(x) \) is strictly increasing, \( g(x) \) is strictly convex and thus, we can write \( g'(b) \) explicitly by solving the first order condition in \( x \). This gives us,

\[
b = g'(x) \iff b = f(x) \iff f^{-1}(b) = x.
\]

Hence, we have that

\[
g'(b) = bf^{-1}(b) - \int_0^{f^{-1}(b)} f(t) \, dt.
\]

The upper integral limit suggests that we consider the change of variable \( f^{-1}(s) \mapsto t \) yielding

\[
g'(b) = bf^{-1}(b) - \int_0^b \frac{s \, ds}{f'(f^{-1}(s))}.
\]

Letting \( u = t \) and \( v = f^{-1}(t) \) and using integration by parts it follows that

\[
g'(b) = \int_0^b u'v \, dt = \int_0^b f^{-1}(t) \, dt.
\]

Young’s Inequality follows immediately. The condition for equality follows from equation (3).

By choosing \( f(t) \) carefully, Young’s inequality can generate the algebraic facts that lay the foundation for three workhorse inequalities in probability theory - the Cauchy-Schwarz, Hölder and Minkowski inequalities - and possibly others. For example, letting \( f(x) = x \) yields

\[
ab \leq \frac{a^2}{2} + \frac{b^2}{2}.
\]
and Cauchy-Schwarz follows soon after. If we let \( f(x) = x^{p-1} \) for \( p > 1 \), Young’s Inequality gives us
\[
ab \leq \frac{a^p}{p} + \frac{b^q}{q}
\]
where \( q = p/(p - 1) \). This algebraic inequality is really the basis for Hölder’s and, subsequently, Minkowski’s inequality.

Young’s inequality can be thought of as something like a practical generator of convex conjugate pairs. We’ve seen how two simple choices for \( f(t) \) can generate the conjugate pairs that form the algebraic inequalities upon which several inequalities in probability theory are based. Understanding the structure of the Legendre-Fenchel transformation gives us some insight into the geometry of these classic inequalities.

3. Cramér-Chernoff Method

We conclude with an example of how the geometry of the Legendre-Fenchel transform can help us understand the behavior of tail probabilities. Let \( X \) be a random variable. Let \( m \) be the log-moment generating function of \( X - E(X) \),
\[
P(X - E(X) \geq t) = P \left( e^{\lambda(X - E(X))} \geq e^{\lambda t} \right) \leq e^{-\lambda t} E \left( e^{\lambda(X - E(X))} \right) = e^{-(\lambda - m(\lambda))}.
\]
By optimizing over \( \lambda \geq 0 \) we get the tail bound \( P(X - E(X) \geq t) \leq e^{-m^c(t)} \).
Bounding tail probabilities in this manner is known as the Cramér-Chernoff Method [see Dembo and Zeitouni, 1998]. It appears that one need not know anything about convex conjugates to arrive at and make use of this bound. Nonetheless, the duality of \( m^c(t) \) and \( m(\lambda) \) can reveal the geometric intuition underlying this bound.

The question that we are interested in is: how fast does \( m^c(t) \) grow in \( t \)? More specifically, at what exponential rate does \( P(X \geq t) \) tend to 0? The following lemma will allow us to examine the shape of either \( m(\lambda) \) or \( m^c(t) \) to answer this question. We often have \( m(\lambda) \) more readily available making this equivalence practically useful when problem solving as long as we know what features of \( m(\lambda) \) to look for.

**Lemma 7.** If \( f \) is convex and differentiable, then \( f \) admits a supporting line at \( x \) with gradient \( k \) if and only if \( f^c \) admits a supporting line at \( k \) with gradient \( x \).

This lemma follows from Lemma 3 and Lemma 5. What this means is that if \( m(\lambda) \) is relatively flat then \( m^c(t) \) will grow relatively fast, leading to a tighter bound on \( P(X \geq t) \). The converse also holds. For example, if \( m(\lambda) \) is such that there is an asymptote at \( t = \alpha \) then Lemma 7 implies that
\[
P(X \geq t) \leq \exp(-\alpha t + O(1)).
\]
A concrete example may help illustrate this relationship:

**Example 8.** Let \( X \sim \chi_2^2 \). Then \( M(\lambda) = (1 - 2\lambda)^{-1/2} \), \( m(\lambda) = -\log(1 - 2\lambda)/2 \) and \( m^c(t) = (t/2) - (1/4) + \log(1/(2t))/2 \). Now observe that \( m(\lambda) \to \infty \) as \( \lambda \to 1/2 \) and that \( \frac{1}{\lambda^2} m^c(t) = (1/2) - 1/(2t) \). From this we see that
\[
P(X \geq t) \leq \exp(-t/2 + O(1))
\]
as expected.
A popular class of random variables that is better behaved than the previous example are sub-Gaussian random variables. Such random variables have logarithmic MGFs that satisfy $m(\lambda) \leq \lambda^2 \sigma^2 / 2$ for some $\sigma^2$ and for all $\lambda$. It follows that,

$$m^c(t) \geq \sup_{\lambda \in \mathbb{R}} \{ t \lambda - \sigma^2 \lambda^2 / 2 \} = \frac{1}{2\sigma^2} t^2;$$

$$\implies P(X \geq E(X) + t) \leq e^{-t^2/2\sigma^2} \quad \text{and} \quad P(X \leq E(X) - t) \leq e^{-t^2/2\sigma^2}.$$

Of course, the standard normal is a member of this family. By Hoeffding’s inequality, a bounded random variable $a \leq X \leq b$ is also sub-Gaussian with $\sigma^2 = (b-a)^2 / 4$.

References


