Legendre transformation and information geometry

CIG-MEMO #2, v1

Frank Nielsen
École Polytechnique
Sony Computer Science Laboratorie, Inc
http://www.informationgeometry.org

September 2010

Abstract

We explain geometrically the Legendre transformation for a strictly convex function $x \in \mathcal{X} \mapsto F(x)$, by first "plotting" its graph, and then reinterpreting this graph as the intersection of its supporting half-spaces. A supporting half-space is parameterized by a dual "slope" parameter $y = \nabla F(x) \in \mathcal{Y}$, and the set of supporting half-spaces yields a convex conjugate function $F^*(y)$ such that $F^*(y) = \max_x\{x^T y - F(x)\}$, maximized for $x = \nabla F(y)$: $F^*(y) = \nabla F(y)^T y - F(\nabla F(y))$. Convex conjugates encode dually the same shape. It follows from the Legendre-Fenchel inequality, a family of non-metric distances $B_{F,F^*}(x, y) = F(x) + F^*(y) - x^T y \geq 0 \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$ that plays the role of canonical divergences of flat spaces in information geometry. Properties of the Legendre transformation are finally briefly listed.

Legendre transformation is at the heart of the duality principle of flat information geometries [1]. Let us explain intuitively this transformation using geometric reasoning. (We shall skip proofs and concentrate on the essence of the transformation instead.) Consider a strictly convex function $F(x)$ for $x \in \mathcal{X}$, and let us plot its graph $\mathcal{F} = \{(x, F(x)) \mid x \in \mathcal{X}\}$. For d-variate functions $F(x_1, ..., x_d)$ with $x = (x_1, ..., x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$, the epigraph is the $(d+1)$-dimensional convex object $\mathcal{O} = \{(x, z) \mid x \in \mathcal{X}, z \geq F(x)\}$. Now let us forget for a while about the x-coordinate system, and "look" at the convex object $\mathcal{O}$ encoding the function. How can we describe (i.e., parameterize) its boundary representation $\partial\mathcal{O}$ (see Figure 1)? Well, we may obviously choose for a *point* $P \in \partial\mathcal{O}$ the x-coordinate system provided by the orthogonal projection: $x(P) = x_P$ and $z_P = F(x_P) = F(x(P))$ so that P has coordinates (x_P, z_P) in the x-coordinate system.

We write $\nabla F(x) = (\frac{\partial F}{\partial x_1}, ..., \frac{\partial F}{\partial x_d})$ for the gradient of F evaluated at x (for univariate function $\nabla F(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$ is the derivative.). Let us consider the *tangent hyperplane* H_P to $\partial\mathcal{O}$ at P of equation $H_P : z = (x - x_P)^T \nabla F(x_P) + F(x_P)$, where $\nabla F(x_P)$ denote the slope parameters of the hyperplane. Let H_P^+ denote the corresponding upper half-space

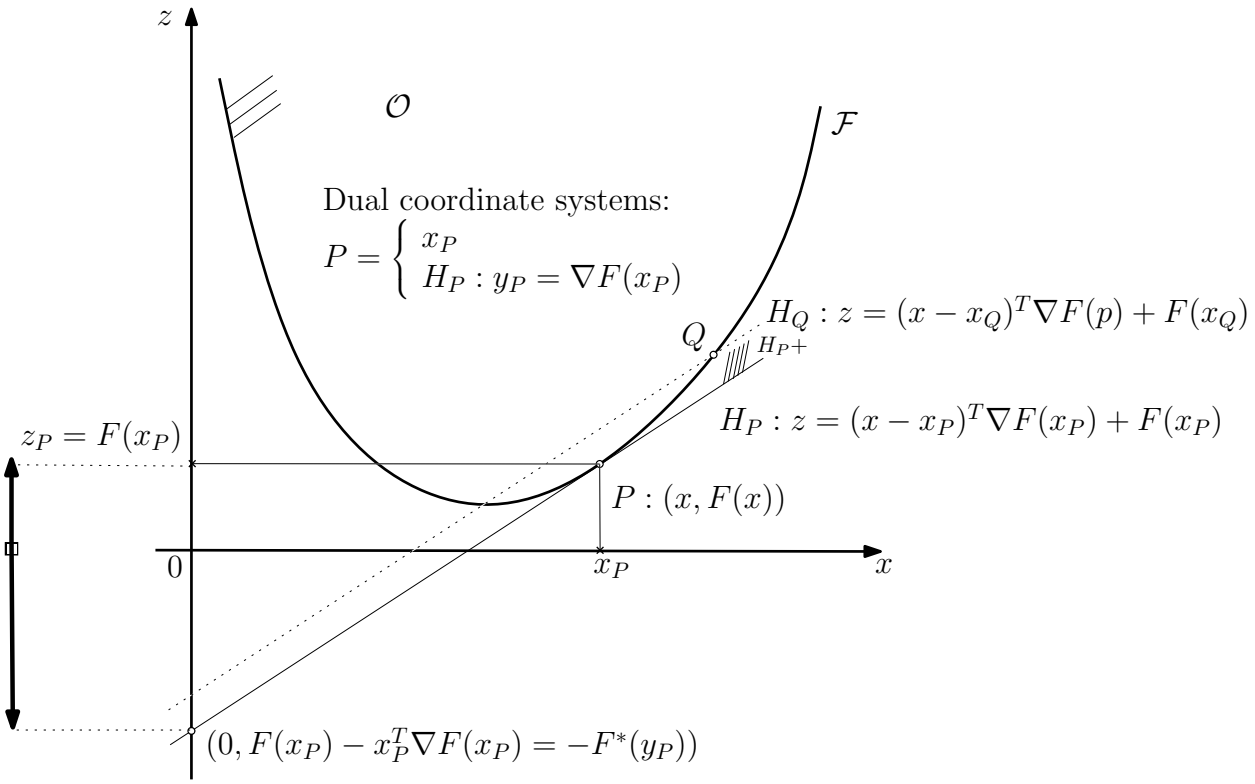

Figure 1: Illustration the Legendre transformation of a strictly convex function: A point P on the boundary of $\mathcal{O}$ can either be parameterized by using the x-coordinate system, or by using the dual slope $y = \nabla F(x)$ coordinate system. For a point $P \in \partial O$ with x-coordinate x_P, and tangent parameter $y_P = \nabla F(x_P)$, the Legendre conjugate $F^*(y)$ reads as the intersection of the hyperplane H_P with the the z-axis. The object $\mathcal{O}$ is either interpreted as the convex hull of the points, or dually as the intersection of the supporting half-spaces.

$z \geq (x - x_P)^T \nabla F(x_P) + F(x_P)$. Now, the key to understand Legendre transformation is to observe that for a "slope" $y = \nabla F(x)$, there is only a *unique* point of $\partial\mathcal{O}$ that admits a tangent hyperplane of that slope. Indeed, since $F(.)$ is strictly convex function, we have $\nabla^2 F(x) \succ 0$ (i.e., positive definite Hessian), and therefore its gradient $\nabla F(x)$ is strictly monotonous increasing: $\forall i \in \{1, ..., d\}, \frac{\partial F}{\partial x_i} \nearrow$. Thus, we may describe as well the boundary of $\mathcal{O}$ using this alternative slope parameter $y = \nabla F(x)$. The point P is also described by the unique point that admits the tangent hyperplane with slope $y = y(P) = \nabla F(x_P)$. Therefore we can identify a point $P \in \partial\mathcal{O}$, either by its x-coordinates or its y-coordinates (with $y = \nabla F(x)$): Namely, we have exhibited a *dual coordinate system*.

How can we write the boundary $\partial\mathcal{O}$ of $\mathcal{O}$ in the y-coordinate system? Any hyperplane of a given slope y passing through a point of $\partial\mathcal{O}$ has a unique point on the z-axis. Indeed, those hyperplanes of fixed slope y are parallel to each others and of generic equation:

$$H_Q : z = (x - x_Q)^T y + F(x_Q) \tag{1}$$

Among those parallel hyperplanes, H_P is the unique hyperplane which minimizes its z-intersection: $z = -x_P^T \nabla F(x_P) + F(x_P)$. That is,

$$P = \arg\min_{Q \in \partial\mathcal{O}} \{-x_Q^T y + F(x_Q)\} = \arg\max_{Q \in \partial\mathcal{O}} \{x_Q^T y - F(x_Q)\} \tag{2}$$

Thus we can read the "function shape" $\partial\mathcal{O}$ using another function G parameterized by the slope $y = \nabla F(x) \in \mathcal{Y}$:

$$G(y) = \max_{x \in \mathcal{X}} \{x^T y - F(x)\}, \tag{3}$$

This defines the Legendre transformation. G is called the *convex conjugate* of F. The right-hand side of Eq. 3 is a strictly concave minimization optimization (sum of an affine term $x^T y$ with a strictly concave function $-F$) with unique maximum x^* found by setting the derivatives to zero: $\nabla_x G(y) = y - \nabla F(x^*) = 0$. That is, $y = \nabla F(x^*)$, the dual parameter $x^* = (\nabla F)^{-1}(y)$ as expected.

In short, the Legendre-Fenchel transformation encodes the "function shape" equivalently in the dual coordinate systems. Let $F^* = G$, then it can be shown that F^* is also strictly convex, and that moreover the conjugation is involutive: $F^{**} = F$. The convex conjugate pair (F, F^*) is related by the functional equality $\nabla F^* = (\nabla F)^{-1}$, or equivalently by $\nabla F = (\nabla F^*)^{-1}$: convex conjugates are reciprocal inverse of each other. Thus a simple rule of thumb for calculating the Legendre-Fenchel transformation consists in first computing the derivate F, then take its functional inverse ∇F, and finally compute the anti-derivative of $(\nabla F)^{-1}$ by integration. We get $F^* = \int (\nabla F)^{-1}$. We can bypass the anti-derivative step by plugging $x^* = \nabla F(y)$ in Eq. 3:

$$F^*(y) = (\nabla F)^{-1}(y)^T y - F((\nabla F)^{-1}(y)), \tag{4}$$

(Unfortunately, we may not always compute the function inverse in closed-form, so that it may be required sometimes to use numerical root solver to approximate F^*.)

Thus we have shown that convex object $\mathcal{O}$ defined as the epigraph of $(x, F(x))$ can also be *equivalently* defined as the intersection of all supporting half-spaces $H_P^+ : z \geq x^T y_P - F^*(y_P)$ with $y_P = \nabla F(x_P)$: $\mathcal{O} = cap_{y_P=\nabla F(x_P)} H_P^+$.

For example, consider $F(x) = x \log x$, Shannon information. We have $F'(x) = 1 + \log x = y$, $F'^{-1}(y) = \exp(y-1) = (F^*)'(y)$, and therefore $F^*(y) = \exp(y-1)$. Observe that domains $\mathcal{X} = \mathbb{R}_*^+$ and $\mathcal{Y} = \mathbb{R}$ do not coincide. Note that minimizing the convex function F amounts to set its gradient to zero: $\min_x F(x) \Rightarrow \nabla F(x) = 0 \Leftrightarrow x = (\nabla F)^{-1}(0) = (\nabla F)^*(0)$. That is, the minimum of a convex optimization problem writes simply as the gradient of the convex conjugate evaluated at zero. For Shannon information, we have $\max_{x \in [0,\infty)} F(x) = x \log x = (F^*)'(0) = e^{-1} \simeq 0.367879...$ We check that $F^*(y) = y \exp(y-1) - \exp(y-1)(y-1) = \exp(y-1)$.

One can check that the Legendre conjugate of $F(x) = \log(1+e^x)$ is $F^*(y) = y \log y + (1-y)\log(1-y)$. Some convex functions are Legendre self-dual: $F(x) = x^2$ or $F(x) = -\log x$ (on $x \in \mathcal{X} = (0, \infty)$).

The strict convexity of the conjugate F^* follows from $\nabla^2 F^*(y) = (\nabla^2 F)^{-1}(x)$ with $x = \nabla F^*(y)$ (the inverse of a positive definite matrix being positive definite). That is, $\nabla^2 F^*(\nabla F(x)) = (\nabla^2 F)^{-1}(x)$.

Legendre transformation is at the very heart of information geometry from the Fenchel-Young inequality. The z-coordinate of the point P, can either be obtained as $z = F(x_P)$ or as $z = y_P^T x_p - F^*(y_P)$ with $y_P = \nabla F(x_P)$. That is, $z = F(x_P) = y_P^T x_P - F^*(y_P)$. This equality becomes an inequality if $y_P \neq \nabla F(x_P)$:

$$F(x) + F^*(y) \geq x^T y \tag{5}$$

which holds with equality if and only if $y = x^* = \nabla F(x)$. This inequality allows to define the *canonical form* of Bregman divergences fully characterizing dually flat spaces [1]:

$$\begin{aligned} \mathcal{X} \times \mathcal{Y} &\rightarrow \mathbb{R}^+ & (6) \\ B_{F,F^*}(x:y) &= F(x) + F^*(y) - x^T y \geq 0. & (7) \end{aligned}$$

The Fenchel-Young inequality can also be easily proved geometrically as follows:

Let $F(x) = \int_0^x \nabla F(t) \mathrm{d}t$ and $F^*(y) = \int_0^y \nabla F^*(t) \mathrm{d}t$ with $\nabla F^* = (\nabla F)^{-1}$. Since both ∇F and $(\nabla F)^{-1}$ are monotonically increasing, we have

$$\begin{aligned} \int_0^x \nabla F(t) \mathrm{d}t + \int_0^y \nabla (\nabla F)^{-1}(t) \mathrm{d}t &\geq \langle x, y \rangle, & (8) \\ F(x) + F^*(y) &\geq \langle x, y \rangle. & (9) \end{aligned}$$

1 Properties

The Legendre-Fenchel transformation enjoys many properties that we concisely list below:

Scaling.

$$F(x) = \lambda G(x) \Rightarrow F^*(y) = \lambda G^*\left(\frac{y}{\lambda}\right), \quad (10)$$

$$F(x) = G(\lambda x) \Rightarrow F^*(y) = G^*\left(\frac{y}{\lambda}\right) \quad (11)$$

Translation.

$$F(x) = G(x) + \lambda \Rightarrow F^*(y) = G^*(y) - \lambda, \quad (12)$$

$$F(x) = G(x + x_0) \Rightarrow F^*(y) = G^*(y) - y^T x_0 \quad (13)$$

Inversion.

$$F(x) = G^{-1}(x) \Rightarrow F^*(y) = -yG^*\left(\frac{1}{y}\right)$$

Infimal convolution. Let the infimal convolution of two functions F and G be defined as

$$(F *_{\inf} G)(x) = \inf\{F(x-t) + G(t) \mid t \in \mathcal{X}\} \quad (14)$$

Then the Legendre conjugate of the infimal convolution of two functions is equal to the elementary Legendre convex conjugates:

$$(F *_{\inf} G)^* = F^* + G^*$$

2 Historical notes

The conjugation transformation is named after French scholar Adrien-Marie Legendre (1752-1833) and in honor of German mathematician Werner Fenchel (1905-1988) for its extension to arbitrary dimensions. The Fenchel-Young inequality originates from its connection to Young inequality: $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ for $a, b, p, q > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$ (with equality for $a^p = b^q$). Indeed, Fenchel inequality $f(a) + f^*(b) \geq ab$ includes Young inequality for $f(a) = a^p/p$ and its Legendre convex conjugate $f^*(b) = b^q/q$ for $\frac{1}{p} + \frac{1}{q} = 1$. William Young (1863-1942) was an English mathematician who made significant contributions to functions of complex variables. Besides information geometry, Legendre transformation plays a fundamental role in formulating problems of thermodynamics, and in Hamilton-Lagrange mechanics.

The Legendre transformation can be extended to arbitrary element *types* using a corresponding inner product:

$$F^(y) = \max_{x \in \mathcal{X}} \langle x, y \rangle - F(x)$$

For example, the inner product of complex matrices is defined as the Hilbert-Schmidt inner product: $\langle X, Y \rangle = \mathrm{Tr}(X^* \times Y)$, where X^* denote the matrix conjugate.

Legendre transformation has also been extended to non-convex functions and non-trivial topology domain $\mathcal{X}$ (eg., $\mathcal{X}$ being a circle).

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

How to cite this document

```
@techreport{cig-memo2,
      author       = "Frank Nielsen",
      title        = "Legendre transformation and information geometry",
      number       = "CIG-MEMO2",
      month        = "September",
      year         = "2010",
      note  = "http://www.informationgeometry.org"
}
```