A graphical derivation of the Legendre transform

Sam Kennerly

April 12, 2011

The Legendre transform is a trick for representing a function in terms of its first derivative. It is simple to define and widely used in physics and applied sciences. Despite its popularity, it is often presented in a haphazard way. While mathematically rigorous descriptions are arguably unnecessary for many applications, some caution is necessary to avoid serious errors in practice. A few common sources of confusion are:

- failing to clearly state the necessary existence/uniqueness conditions,
- using notation which confuses numbers with functions, and
- misinterpreting the somewhat-ambiguous formula $px - f(x)$.

The author has committed each of these errors. The second error is especially popular with physicists. The symbol $y(x)$ is often used to represent both "the function $y()$" and "the value of y at x." This abuse of notation is usually harmless, but it can be dangerous when change-of-variable techniques are used. Here I will use y to mean a number, $y()$ to mean a function, and $y(x)$ to mean "the output of $y()$ when given x as an input."

After completing nearly all of this article, I discovered a recently-published paper called *Making Sense of the Legendre Transform* which presents many of the same ideas. I found it generally clear and helpful, so I have listed it as a reference.[1]

1 Existence/uniqueness conditions for a Legendre transform

Suppose all of the following statements are true:

1. A well-behaved function $f()$ is defined over some chunk D of the real line.
2. For any $x \in D$, you know how to find $f(x)$ and $\frac{d}{dx}f(x)$, which I will call $f'(x)$.
3. The graph of $f(x)$ always curves upward: for any $x \in D$, $f''(x) > 0$.[2]

[2] If $f''(x) < 0$ always, then define a new function $\tilde{f}() = -f()$ and Legendre transform $\tilde{f}()$ instead.

Condition 1 is deliberately vague; what do "well-behaved" and "chunk" mean? The point is: when using functions that fail common tests (e.g. continuity, non-singularity, smoothness), be careful. A more rigorous treatment than the one provided here may be necessary.

Condition 2 simply requires that an explicit formula for $f(x)$ is known, its derivative can be found either by hand or by computer, and that derivative is also well-behaved.

Condition 3 is not always stated explicitly, but it should be. Legendre transformations behave very badly if the curvature of $f()$ changes sign as x changes. (If $f''(x)$ fails to exist at some points, see the subsection "Convex functions and convex sets.")

Suppose that instead of using x as a variable, you would prefer a new variable p such that $p(x) = f'(x)$. The Legendre transform produces a formula, in terms of p, for a new function $g()$. The transform is invertible, so knowing $g(p)$ tells you everything about $f(x)$.

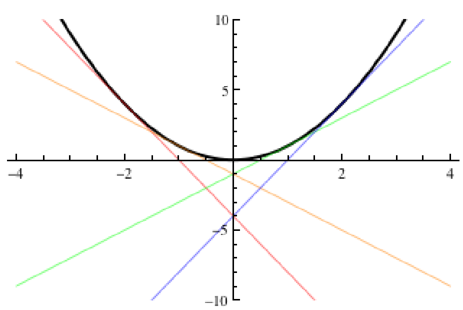
2 Geometric interpretation of the Legendre transform

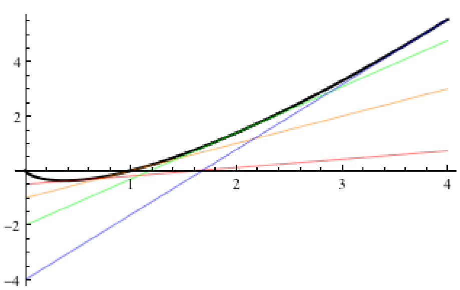
Plot $f(x)$. At each point, imagine a line tangent to the plot. This line intersects $(x, f(x))$ and has slope $p = f'(x)$. Any straight line with slope p must look like this for some $g \in \mathbb{R}$:

$$y(x) = px - g$$

Here g means "the negative y-intercept of the line tangent to $f()$ at the point $(x, f(x))$." (We could have defined g to be the *positive* y-intercept, but that's not the usual convention.)

Since $f''(x) > 0$ everywhere, there is only one tangent line for each possible slope p. Draw pictures to convince yourself that if a function always curves upward, it can't have two tangent lines with the same slope. Here are two examples: $f(x) = x^2$ and $f(x) = x\log(x)$. For each plot, $f(x)$ is in black and several tangent lines to $f()$ are shown in color.[3]





For each possible slope p, there is exactly one tangent line. That tangent line has its y-intercept at $y = -g(p)$. We want to find the function $g()$ that maps p's to g's.

The really useful thing about $g()$ is this: each point $(x, f(x))$, has exactly one "evil twin" point $(p, g(p))$. Knowing $g()$ then gives us complete information about the $f()$ and vice versa. (See the subsection "Convex functions and convex sets" for weird exceptions.)

[3] $f(x) = x\log(x)$ behaves badly when $x \leq 0$, so I've shown it only on the domain $D = (0, \infty)$.

Given a slope p, define $x(p)$ to be the value of x such that $f'(x) = p$.[4] The negative-y-intercept of the tangent line with slope p is found by setting $px - g = f(x)$.

The recipe for the Legendre transform is:

0. Check that $f()$ satisfies the existence/uniqueness conditions.
1. Define a new function $p()$ such that $p(x) = f'(x)$. Invert $p()$ and call the result $x()$.
2. Define g to be the negative of the y-intercept of the line tangent to $f()$ at x:

$$g = p(x) \cdot x - f(x)$$

3. Use the formula for $x(p)$ to write the x's as functions of p. Call the result $g(p)$.

$$g(p) = p \cdot x(p) - f(x(p))$$

Just be careful to remember what $x(p)$ means: it is the value of x at which the slope of $f'()$ is $f'(x) = p$. Otherwise this equation won't make any sense.

3 Examples of Legendre transforms

3.1 Example #1: $f(x) = x^2$ with $x \in \mathbb{R}$

0. $f()$ is well-behaved, $f'()$ is well-behaved, and $f''() > 0$.

1. Define $p(x) = f'(x) = 2x$. Invert this to find $x(p)$: $\quad p = 2x \Leftrightarrow x = \frac{1}{2}p$

2. Define $g = p(x) \cdot x - f(x)$: $\quad g = 2x \cdot x - x^2 = x^2$

3. Use $x(p) = \frac{1}{2}p$ to write the x's as functions of p: $\quad g(p) = (\frac{1}{2}p)^2 = \frac{1}{4}p^2$

The Legendre transform is $f(x) = x^2 \;\Leftrightarrow\; g(p) = \frac{1}{4}p^2$.

3.2 Example #2: $f(x) = x\log(x)$ with $x \in \mathbb{R}$ and $x > 0$

0. If $x > 0$, then $f()$ is well-behaved, $f'()$ is well-behaved, and $f''() > 0$.

1. Define $p(x) = f'(x) = \log(x) + 1$. Invert this to find $x(p)$.

$$p = \log(x) + 1 \;\Leftrightarrow\; \log(x) = p - 1 \;\Leftrightarrow\; x = e^{p-1}$$

2. Define $g = p(x) \cdot x - f(x)$.

$$g = (\log(x) + 1)x - x\log(x) = x$$

3. Use $x(p) = e^{p-1}$ to write the x's as functions of p.

$$g(p) = e^{p-1}$$

The Legendre transform is $f(x) = x\log(x) \;\Leftrightarrow\; g(p) = e^{p-1}$.

[4] Equivalently, $x()$ is the inverse function of $f'()$. The domain of $x()$ is $S = \{$ all possible slopes $p\}$. See the subsection "Legendre-transforms as inverse-derivative pairs" for details.

3.3 Physics example: Hamiltonian of a $360°$ pendulum

Imagine a pendulum made of a very light, rigid rod of length R with a dense, point-like blob of mass m on one end. The other end is attached to a ball bearing which allows the pendulum to rotate 360° in a vertical plane. Define θ to be the angle between the rod and a vertical line and set $\theta = 0$ when the blob is at maximum height. Choose positive θ to be clockwise or counter-clockwise.[5] Ignore friction but don't ignore gravity.

The (approximate) gravitational potential energy of this object is $V = mgy = mgR(\cos\theta)$. The (approximate) rotational kinetic energy is $K = \frac{1}{2}I\omega^2 = \frac{1}{2}mR^2\omega^2$, where ω is the pendulum's angular velocity. The Lagrangian describing the system is $\mathcal{L} = K - V$.

$$\mathcal{L}(\theta, \omega) = K - V = \frac{1}{2}mR^2\omega^2 - mgR(\cos\theta)$$

The Hamiltonian of this system is found by Legendre-transforming $\mathcal{L}$ to remove the variable ω. (The variable θ comes along for the ride. For our purposes, θ can be thought of as a constant during the Legendre-transform process.) First, define $p(\omega) = \mathcal{L}'(\omega)$:

$$p(\omega) = \mathcal{L}'(\omega) = mR^2\omega$$

Is $\mathcal{L}''(\omega) > 0$ for all ω? Since $\mathcal{L}''(\omega) = mR^2$ and $m > 0$, it is. Note that p has a physical interpretation as the pendulum's angular momentum $mR^2\omega = I\omega$.

Now invert $p(\omega)$ to find $\omega(p) = \frac{p}{mR^2}$. Define $g = p(\omega)\cdot\omega - \mathcal{L}(\omega)$ as usual, use $\omega(p)$ to write everything in terms of p's, and call the result $g(p)$:

$$g(p) = \frac{p^2}{mR^2} - \frac{1}{2}mR^2\left(\frac{p}{mR^2}\right)^2 + mgR(\cos\theta) = \frac{p^2}{2mR^2} + mgR(\cos\theta)$$

Remembering that θ is not really a constant, we should call it $g(\theta, p)$. Also, traditional notation uses H and L instead of g and p for "Hamiltonian" and "angular momentum."

$$H(\theta, L) = \frac{L^2}{2mR^2} + mgR(\cos\theta)$$

This is the pendulum's Hamiltonian function. It has a physical interpretation as the total (kinetic + potential) energy of the pendulum in terms of angular position and momentum.

In most simple physical systems like this one, using a Legendre transform to find the particle's Hamiltonian seems like extra work for no clear benefit; why not just write $H = K + V$ in the first place? For many practical calculations, this is an excellent criticism. The method is primarily important for providing a theoretical motivation for quantum mechanics.[6]

---

[5] I choose these coordinates so that the Lagrangian does not depend on the sign of θ.

[6] If you want to sound very technical, tell people "the Legendre transform of a Lagrangian function constructs an invertible map between solutions of an n-dimensional second-order equation of motion and a flow on a $2n$-dimensional symplectic manifold," or something like that.

3.4 Thermodynamic potentials

A thermodynamic system can be completely described by its internal energy function $U(S,V,N)$. Here U is internal energy, S is entropy, V is volume, and N is particle number. The partial derivatives of $U(S,V,N)$ have names suggestive of their usual physical interpretations. They are neatly summarized in the **fundamental thermodynamic relation**:

$$dU = TdS - PdV + \mu dN$$

where T, P, and μ are called temperature, pressure, and chemical potential.

One problem with this description is that entropomometers, if such things exist, are hard to find. Temperature-measuring devices are much more convenient. We'd like to use T as a variable instead of S, but first we should check if $\frac{\partial^2}{\partial S^2}U(S,V,N) > 0$.

$$T = \left(\frac{\partial U}{\partial S}\right)_{V,N} \qquad \left(\frac{\partial^2 U}{\partial S^2}\right)_{V,N} = \left(\frac{\partial T}{\partial S}\right)_{V,N} = \left(\frac{\partial S}{\partial T}\right)_{V,N}^{-1} > 0$$

which is the physically plausible claim that raising the temperature of a thermodynamic system while holding everything else but U constant will increase its entropy.

Once $T(S)$ is found, invert it to find $S(T)$. Legendre-transform $U(S,V,N)$ to produce

$$TS(T) - U(S(T),V,N)$$

The **Helmholtz energy** $A(T,V,N)$ of a system is conventionally defined to be -1 times the Legendre transform of $U(S,V,N)$ with S removed in favor of $T = \partial U/\partial S$.[7]

$$A(T,V,N) = U(S(T),V,N) - T\cdot S(T)$$

The other variables V, N can be Legendre-transformed as well. Suppose we don't mind S, but V annoys us and we prefer to use P as a variable in our thermodynamic potential. We'll need to check that $U(S,V,N)$ is concave-up for V:

$$P = -\left(\frac{\partial U}{\partial V}\right)_{S,N} \qquad \left(\frac{\partial^2 U}{\partial V^2}\right)_{S,N} = -\left(\frac{\partial P}{\partial V}\right)_{S,N} > 0$$

which another physically plausible claim: increasing volume *decreases* pressure, all other things being equal. Inverting this to find $V(P)$, we can now define

$$H(S,P,N) = U(S,V(P),N) + P\cdot V(P)$$

The + before $P\cdot V(P)$ is a side effect of defining P with the wrong sign and multiplying everything by an overall factor of -1.[8] (By convention, the **enthalpy** $H(S,P,N)$ is actually -1 times the V-Legendre transform of $U(S,V,N)$.)

---

[7] A is for *Arbeit*, which is German for "work." The letter F, for "Free energy," is also popularly used.

[8] We could have defined $P = \frac{\partial U}{\partial V}$, but that would be the negative of what we intuitively associate with the word "pressure." Perhaps it should be called "vacuosity."

If we can S-Legendre-transform U and V-Legendre-transform U, can we do both? Yes, and the result, including the usual minus sign conventions, is called the **Gibbs free energy**.

$$G(T, P, N) = U(S(T), V(P), N) - T \cdot S(T) + P \cdot V(P)$$

We can continue in this fashion by replacing N with μ to form the **Landau potential** (a.k.a. **grand potential** or Φ **potential**). A relatively easy-to-remember shorthand is:

$$A = U - TS \qquad H = U + PV \qquad G = U - TS + PV \qquad \Phi = U - TS + PV - \mu N$$

Performing these transformations for realistic internal energy functions can be rather time-consuming and mistake-prone, which is why I have not included explicit examples.

4 Technical details

4.1 Legendre transforms as inverse-derivative pairs

Another way to define the Legendre transform is "the function whose derivative is the inverse function of $f'()$." This definition is not very geometric, but it suggests a useful idea:

To "undo" a Legendre transform $f(x) \to g(p)$, Legendre transform $g(p)$ again to get $f(x)$.

More precisely, the Legendre transform can be thought of as an operator $\hat{L}$ that maps functions to other functions: $\hat{L}f() = g()$. It turns out to be its own inverse operator: $\hat{L}(\hat{L}f()) = \hat{L}^2 f() = \hat{1}f() = f()$. The invertibility of $\hat{L}$ is essential to its value as a method for changing representations without destroying information. I won't rigorously prove these claims, but try twice-Legendre-transforming some of the examples to get the idea.

For this new inverse-derivative definition to make any sense, the inverse function of $f'()$ must exist. First, assume $f'()$ exists on a domain D. (If not, there's no purpose attempting a Legendre transform!) Define the **slope set** S of $f()$ as the range of $f'()$:

$$S = \{\ p \mid f'(x) = p \text{ for some } x \in D\}$$

S is the set of all possible slopes of tangent lines to $f()$. By definition, $f'()$ is onto S. If $f'()$ is also one-to-one on D, then $f'() : D \to S$ is invertible. If $f'()$ is not one-to-one, then "the inverse function of $f'()$" will be ambiguous.

We assume $f''(x) > 0$ for all $x \in D$, so $f'()$ is monotonically increasing on D, so $f'()$ must be one-to-one.[9] Then there exists $x() : S \to D$ which outputs exactly one x-value in D for each slope $p \in S$. The Legendre transform $\hat{L}f() = g()$ is the unique function such that:

1. $g'()$ and $f'()$ are each other's inverse function: $g'(f'(x)) = x$ and $f'(g'(p)) = p$.
2. $f(x) + g(p(x)) = xp(x)$ for all $x \in D$.

[9] The subsection "Convex functions and convex sets" covers exceptions for when $f''()$ does not exist.

To show that this new definition is equivalent to the original, define $p(x) = f'(x)$ and $x(p)$ in the same way as before. Find $g'(p)$ by using the chain rule:

$$g'(p) = \frac{d}{dp}\Big[p \cdot x(p) - f\Big(x(p)\Big)\Big] = x(p) + p \cdot x'(p) - f'\Big(x(p)\Big) \cdot x'(p)$$

By definition $p(x) = f'(x)$, so $f'(x(p))$ is just $p(x(p)) = p$. The only term in $g'(p)$ that does not cancel is $x(p)$, so $g'(p) = x(p)$. This means $g'()$ must be the inverse of $p() = f'()$.

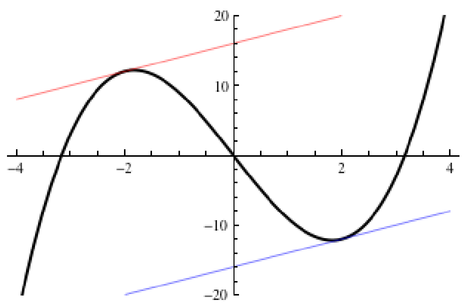
I've assumed that $x'(p)$ exists. To check, use the reciprocity relation $\frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1}$:

$$x'(p) = \Big|_p \frac{dx}{dp} = \Big|_{x(p)} \left(\frac{dp}{dx}\right)^{-1} = \Big(f''(x(p))\Big)^{-1}$$

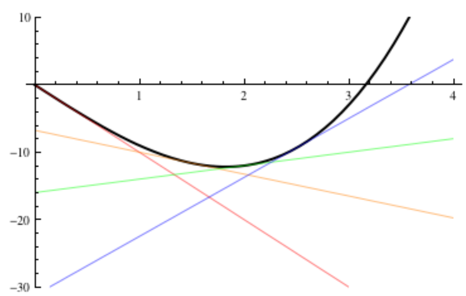
which is bad if $f''(x) = 0$ for some $x \in D$, but I've assumed that $f''(x) > 0$.

Defining $g()$ as "the function whose derivative is the inverse function of $f'()$" is not as geometrically-motivated as the previous tangent-line-intercept definition. However, it does draw attention to an important geometric feature of Legendre transforms.

Specifically, the condition "$f'()$ must be monotonically increasing" is equivalent to "$f()$ cannot have two different points with the same slope."[10] Two examples are shown below:



Left: $f(x) = x^3 - 10x$ with domain $\mathbb{R}$.



Right: $f(x) = x^3 - 10x$ restricted to $\mathbb{R} > 0$.

If $f(x) = x^3 - 10x$, then $f'(x) = 3x^2 - 10$ and $f''(x) = 6x$. This function has $f''(x) \leq 0$ for any $x \leq 0$. The difficulty is shown graphically: the red and blue tangent lines at $x = \pm 2$ have the same slope $p = 2$. Consequently $x(p)$ is ambiguous: does $x(2) = 2$ or does $x(2) = -2$?

Graphically, $f()$ changes curvature from negative to positive as it passes through $x = 0$. Imagine a particle moving from left to right and visualize a tangent line to $f()$ at the location of the particle. The tangent line's slope p goes from positive to negative, then "changes its mind" and becomes positive again. Continuity of $f'()$ forces the line to repeat previous p-values, which ruins the invertibility of $f'()$.

The picture on the right restricts the domain of $f()$ to positive numbers. This restricted version of $f()$ has positive curvature everywhere, so no two tangent lines have the same slope. Consequently $p() = f'()$ is monotonically increasing, its inverse function $x()$ can be defined, and a Legendre transform can be safely performed.

[10] A monotonically decreasing $f'()$ is also acceptable. An equivalent condition is $f''() < 0$ for all $x \in D$,

4.2 Convex functions and convex sets

A function $f()$ is **convex** if for all $w_1 \in [0, 1]$ and $w_2 = (1 - w_1)$,

$$f(w_1 x_1 + w_2 x_2) \le w_1 f(x_1) + w_2 f(x_2)$$

In words, $f()$ of a weighted average of x_1, x_2 is, at most, a weighted average of $f(x_1)$ and $f(x_2)$ using the same weights. I prefer to remember the definition this way:

$$f(\text{weighted average of } x\text{'s}) \le \text{weighted average of } f(x)\text{'s}$$

A **convex combination** of n numbers $x_1, x_2, \ldots, x_n$ is any number of the form:

$$w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \quad \text{such that all } w_k \ge 0 \text{ and } \sum w_k = 1$$

A linear combination is convex if its coefficients are non-negative and sum to 1. Convex combinations can be thought of as weighted averages with weights $\{w_k\}$.

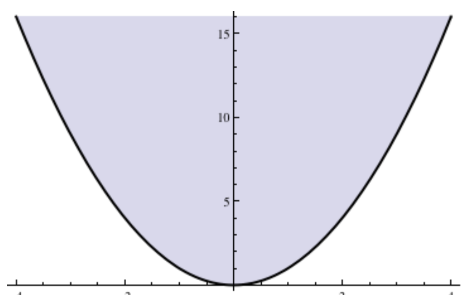
As an example, consider investing money in one of two bank accounts. Each will return $f(r, t) = e^{rt}$ times your investment after t days, but the interest rates r_1, r_2 are random real numbers. Another bank offers to figure out these rates and pay you $e^{\bar{r}t}$ where $\bar{r}$ is the average rate $\frac{1}{2}(r_1 + r_2)$. To keep the arithmetic simple, we'll consider only two strategies:

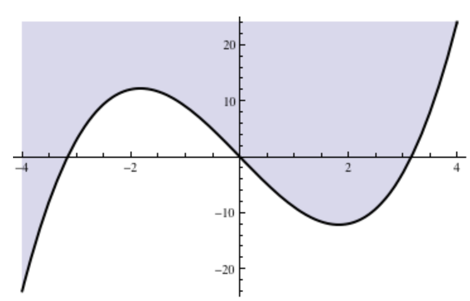
1. Deposit all your money in the $\bar{r}$ account.
2. Deposit half your money in the r_1 account and half in the r_2 account.

The first strategy pays $A_1(t) = e^{\frac{1}{2}(r_1+r_2)t}$ and the second strategy pays $A_2(t) = \frac{1}{2}(e^{r_1 t} + e^{r_2 t})$. Because $f(r, t)$ is convex in r, the first strategy is *never* better. Here's a quick proof:

$$A_2(t) - A_1(t) = \frac{1}{2}(e^{r_1 t} + e^{r_2 t}) - e^{\frac{1}{2}(r_1+r_2)t} = \frac{1}{2}\left(e^{\frac{1}{2}r_1 t} - e^{\frac{1}{2}r_2 t}\right)^2 \ge 0$$

Convex functions are related to **convex sets**. Given a function $f()$, define F_{top} to be all the points "above" $f()$: $F_{top} = \{(x, y) \mid y \ge f(x)\}$. F_{top} is convex if and only if $f()$ is convex. A set S is convex if for any two points $\mathbf{r}_1, \mathbf{r}_2 \in S$, the line segment connecting $\mathbf{r}_1$ to $\mathbf{r}_2$ is contained within S. Loosely speaking, this means convex sets have no dents in them.





The shaded areas above show F_{top} for $f(x) = x^2$ on the left and $f(x) = x^3 - 10x$ on the right. $f(x) = x^3 - 10x$ is not convex, and its F_{top} has a big dent in it. If I choose e.g. $(-4, 0)$ and $(2, 0)$ as $\mathbf{r}_1$ and $\mathbf{r}_2$, the line connecting them strays outside the shaded region.

A Legendre transform can be defined for any convex function in the following way:

$$g(p) = \max_x [px - f(x)]$$

where $\max_x[]$ means "maximum possible value as x varies but p is held constant." The domain of p must be restricted to values such that $g(p)$ is finite; $g()$ is undefined elsewhere.

If $f()$ is differentiable on its domain D, we can find $\max_x[px - f(x)]$ by taking the partial derivative $\frac{\partial}{\partial x}(px - f(x))$ and setting it equal to zero: $p - f'(x) = 0$. Since $f()$ is convex, we know that $p()$ has an inverse function $x()$. That leads to our original definition:

$$g(p) = p \cdot x(p) - f(x(p)) \qquad p(x) = f'(x)$$

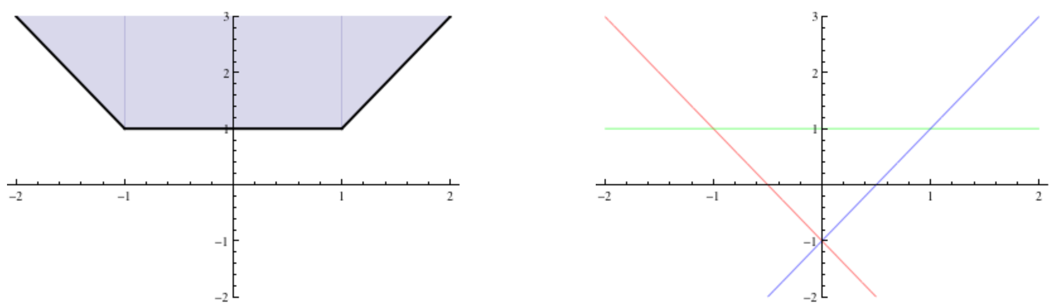
If $f''(x) > 0$, then $f()$ is convex. The converse is not true because convex functions are not necessarily differentiable. A convex function defined on an open interval of $\mathbb{R}$ is always continuous and **almost differentiable** on that interval, which means the number of points at which $f'(x)$ fails to exist is either finite or countably infinite. For example,

$$\text{Bucket}(x) = \{-2x - 1 \text{ if } x < -1,\ 1 \text{ if } -1 \leq x \leq 1,\ 2x - 1 \text{ if } x > 1\}$$

Bucket() is continuous and almost differentiable. (The derivative fails at exactly two points, $x = \pm 1$.) Its slope set has only three elements: $S = \{-2, 0, 2\}$. The Legendre transform of Bucket() need only be defined for $p \in S$. At these p-values, $\max_x[px - f(x)]$ is

$$g(p) = \max_x[px - f(x)] = \{1 \text{ if } p = \pm 2,\ -1 \text{ if } p = 0\}$$

Geometrically, this means "$f()$ has slopes $p \in \{-2, 0, 2\}$. When $p = 2$ or -2, the negative y-intercept of the tangent line is 1. When $p = 0$, the negative y-intercept is -1."



Bucket() and its F_{top} set are shown on the left with its tangent lines on the right. At $x = \pm 1$, there are no tangent lines. However, continuity of Bucket() tells us its values at these "bad" points: Bucket(-1) = Bucket(1) = 1. Any convex function can be re-constructed from its tangent lines in a similar way, which shows the central idea of Legendre transforms:

All information about a convex function is stored in the y-intercepts of its tangent lines.

References

[1] R. K. P. Zia, Edward F. Redish, and Susan R. McKay. Making sense of the legendre transform. *American Journal of Physics*, 77(7):614, 2009.