An Application of Duality

Jason d'Eon

January 30, 2023

Since learning some concepts on duality in convex optimization back in 2019, I have been wanting to come up with a canonical problem that demonstrates the usefulness of strong duality. In linear programming, if the problem has few variables and many constraints, solving the dual can be advantageous in terms of efficiency, but I was looking for an non-linear example where duality can provide insights into the problem. Below is described such a problem where the initial constraint is difficult to optimize around, and computing the dual yields a straightforward convex optimization problem.

1 Notation and Definitions

The problem revolves around probability distributions over a discrete variable x, which we denote with $P(x)$ and $Q(x)$. The expected value of a function $z = f(x)$ is given by $\mathbb{E}[z] = \sum_x z \cdot P(x)$. The Pearson χ^2-divergence $D_{\chi^2}(Q \parallel P)$ is a function of two distributions indicating a level of similarity between them, and it is given by:

$$D_{\chi^2}(Q \parallel P) = \sum_x P(x) \left(\frac{Q(x)}{P(x)} - 1 \right)^2.$$

Additionally, this satisfies that $D_{\chi^2}(Q \parallel P) \geq 0.$ and $D_{\chi^2}(Q \parallel P) = 0$ if and only if $Q(x) = P(x)$ for all x.

2 Problem Statement

Consider a fair 6-sided die, and let $P(x)$ represent the uniform distribution of the outcome of the die, x. Clearly $\mathbb{E}[x] = 3.5$. Now suppose we want to create an unfair die with a larger expected value, without modifying the distribution too much. To this end, we wish to solve the following:

$$\begin{aligned} \max_Q \quad & \mathbb{E}_Q[x] \\ \text{s.t.} \quad & D_{\chi^2}(Q \parallel P) \leq r \end{aligned} \tag{1}$$

where Q is any distribution over the die. In other words, find the distribution in a χ^2-ball of radius r centered around P that maximizes the expected value of the die.

3 Background

3.1 Strong Duality

The main idea behind duality is that for every maximization problem (which we refer to as the *primal*), there is a closely related minimization problem (referred to as the *dual*). In the general case of concave/convex optimization problems, the primal is given by:

$$\begin{aligned} \max_{x} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, n \\ & h_j(x) = 0, \quad j = 1, \ldots, m \end{aligned} \tag{2}$$

where f, g_i, h_i are all concave functions. The Lagrangian function is as follows:

$$\mathcal{L}(x, \lambda, \eta) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \eta_j h_j(x) \tag{3}$$

where λ_i and η_j are the introduced dual variables (very much related to Lagrange multipliers, though in this case we can have inequality constraints). The dual objective function is then given by:

$$g(\lambda, \eta) := \inf_{x} \mathcal{L}(x, \lambda, \eta) \tag{4}$$

which is not to be confused with the g_i constraints. In this case where we started with maximizing a concave function, $g(\lambda, \eta)$ is convex. So the dual problem in full is:

$$\begin{aligned} \min_{\lambda, \eta} \quad & g(\lambda, \eta) \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, n \\ & \eta_j \in \mathbb{R}, \quad j = 1, \ldots, m \end{aligned} \tag{5}$$

It is always the case that the maximizer x^* of the primal and the minimizers λ^*, η^* of the dual will satisfy a condition known as *weak duality*:

$$f(x^*) \leq g(\lambda^*, \eta^*) \tag{6}$$

however, if certain regularity conditions are met, then the solutions will match, which is known as *strong duality*:

$$f(x^*) = g(\lambda^*, \eta^*) \tag{7}$$

Strong duality is also equivalent to asking whether or not a saddle point of $\mathcal{L}(x, \lambda, \eta)$ exists that would simultaneously solve the maximization and minimization problems.

The regularity conditions for strong duality are commonly called constraint qualifications. An example of such a constraint qualification is Slater's condition which for convex optimization problems says that the feasible region has a non-empty interior. In other words, such a condition being met implies that strong duality holds. This is also closely related to what are known as the Karush-Kuhn-Tucker (KKT) conditions. Note that the original problem we stated does satisfy the conditions for strong duality, if $r > 0$.

3.2 Fenchel Conjugate

The Fenchel conjugate (also known as the convex conjugate) is also closely related to duality. For simplicity we will only consider functions of a single variable since it is all we need to solve our original problem, but this can be generalized to functions on any real topological vector space.

Let $f : \mathbb{R} \to \mathbb{R}$. The *Fenchel conjugate* is defined by:

$$f^*(y) := \sup\{yx - f(x)\} \tag{8}$$

It can be shown that f^* must be convex and that $f^{**} \leq f$. In fact, the conditions for strong duality are essentially equivalent to the conditions for $f^{**} = f$ (see the *Fenchel-Moreau theorem*). As an example, equality holds if f is continuous and convex.

The geometric interpretation behind this concept is based around the fact that one can encode a function as the set of hyperplanes that support the convex hull of the epigraph of the function (where the epigraph is the set of points above the function). In fact, this is simply due to the fact that any convex set can be reconstructed from the set of hyperplanes supporting it, which is known as the *supporting hyperplane theorem*. It is typically proven using the *hyperplane separation theorem* (or *Hahn-Banach theorem*).

In one dimension, if f is convex, the definition simplifies down to asking, *"For a given slope of y, how far do you have to translate the line yx so that it is a tangent line of f?"* The value of such a perturbation is $f^*(y)$. If f is not convex, then yx may not be translated to be tangent to f, but rather the convex hull of the epigraph of f.

Example 1. *As an example, take $f = (x-1)^2$, which is convex. To find f^*, we can compute the value of x which achieves the* sup*:*

$$\begin{aligned}\frac{\partial}{\partial x}\left[yx - (x-1)^2\right] &= y - 2(x-1) \\ 0 &= y - 2x + 2 \\ x &= \frac{1}{2}y + 1\end{aligned}$$

then plug it back into the original definition:

$$\begin{aligned}f^*(y) &= \sup\{yx - f(x)\} \\ &= y\left(\frac{1}{2}y + 1\right) - \left(\frac{1}{2}y + 1 - 1\right)^2 \\ &= \frac{1}{4}y^2 + y\end{aligned}$$

Therefore the Fenchel conjugate of $(x-1)^2$ is $\frac{1}{4}y^2 + y$.

In addition, one can see that when we set $y = 1$, we have that $yx - f^*(y) = x - \frac{5}{4}$ is a tangent line of f with a slope of $y = 1$.

Following the same procedure, one can also show that $f^{**}(x) = (x-1)^2$.

It is also fairly straightforward to show that for any scalar c, we have $(cf)^*(y) = c \cdot f^*(y/c)$.

3.3 f-Divergences

The χ^2-divergence belongs to a larger class of divergences known as f-divergences. These are divergences of the form:

$$D_f(Q \parallel P) := \sum_x P(x) f\left(\frac{Q(x)}{P(x)}\right) \tag{9}$$

where f is a convex function, $f(1) = 0$, $f(x)$ is finite on $x > 0$, and $f(0) = \lim_{x \to 0^+} f(x)$.

Some examples from this class include:

$$\begin{aligned} \chi^2\text{-divergence:} &\quad f(x) = (x-1)^2 \\ \text{KL-divergence:} &\quad f(x) = x \ln x \\ \text{reverse KL-divergence:} &\quad f(x) = -\ln x \\ \text{Total variation distance:} &\quad f(x) = \frac{1}{2}|x-1| \\ \text{squared Hellinger distance:} &\quad f(x) = (\sqrt{x}-1)^2 \end{aligned} \tag{10}$$

This class satisfies many nice properties. For example, if $g(x) = f(x) + c(x-1)$ for some $c \in \mathbb{R}$, then $D_g = D_f$. That is, f-divergences are not quite uniquely defined by their generating function f. Therefore, the χ^2-divergence would be equivalently defined using $f(x) = x^2 - 1$ or $f(x) = x^2 - x$.

It is clear that in general divergences are not symmetric, that is, $D_f(Q \parallel P) \neq D_f(P \parallel Q)$. However, if we let $g(x) = xf(1/x)$ for some generating function f, then we get that $D_f(Q \parallel P) = D_g(P \parallel Q)$. Such a transformation of f is called a convex inversion.

4 Problem Solution

Let $f(x) = (x-1)^2$, so that $D_{\chi^2} = D_f$. The primal problem (1) restated is:

$$\begin{aligned} \max_Q \quad & \sum_x xQ(x) \\ \text{s.t.} \quad & D_f(Q \parallel P) \leq r \\ & \sum_x Q(x) = 1 \end{aligned} \tag{11}$$

The Lagrangian dual function is given by:

$$L(Q, \lambda, \eta) = \left(\sum_x xQ(x)\right) + \lambda\left(r - \sum_x P(x) f\left(\frac{Q(x)}{P(x)}\right)\right) + \eta\left(1 - \sum_x Q(x)\right) \tag{12}$$

So if we set $g(\lambda, \eta) = \max_Q L(Q, \lambda, \eta)$, then the dual is:

$$\begin{aligned} \min_{\lambda,\eta} \quad & g(\lambda, \eta) \\ \text{s.t.} \quad & \lambda \geq 0 \\ & \eta \in \mathbb{R} \end{aligned} \tag{13}$$

Let us investigate the dual problem in further detail. We have that:

$$\begin{aligned}
g(\lambda,\eta) &= \eta + \lambda r + \max_Q \sum_x \left(xQ(x) - \eta Q(x) - \lambda P(x) f\left(\frac{Q(x)}{P(x)}\right)\right) \\
&= \eta + \lambda r + \sum_x \max_Q \left(xQ(x) - \eta Q(x) - \lambda P(x) f\left(\frac{Q(x)}{P(x)}\right)\right) \\
&= \eta + \lambda r + \sum_x P(x) \max_{t\geq 0}\{t(x-\eta) - \lambda f(t)\}, \quad t := Q(x)/P(x) \\
&= \eta + \lambda r + \sum_x P(x)(\lambda f)^*(x-\eta) \\
&= \eta + \lambda r + \sum_x P(x)\lambda f^*\left(\frac{x-\eta}{\lambda}\right)
\end{aligned}$$

Since strong duality holds for this problem, we have proven that:

$$\max_Q \{\mathbb{E}_Q[x] : D_f(Q \parallel P) \leq r\} = \min_{\lambda\geq 0,\eta} \left\{ \mathbb{E}_P\left[\lambda f^*\left(\frac{x-\eta}{\lambda}\right)\right] + \lambda r + \eta \right\} \tag{14}$$

We can actually simplify this even further by eliminating λ. Note that $f^*(y) = \frac{1}{4}y^2 + y = (\frac{1}{2}y+1)^2 - 1$. Then:

$$\begin{aligned}
\lambda\,\mathbb{E}_P\left[f^*\left(\frac{x-\eta}{\lambda}\right)\right] + \lambda r + \eta &= \lambda\,\mathbb{E}_P\left[\left(\frac{1}{2}\left(\frac{x-\eta}{\lambda}\right)+1\right)^2 - 1\right] + \lambda r + \eta \\
&= \lambda\,\mathbb{E}_P\left[\frac{1}{4\lambda^2}(x-\eta)^2 + \frac{1}{\lambda}(x-\eta) + 1\right] + \lambda(r-1) + \eta \\
&= \frac{1}{4\lambda}\,\mathbb{E}_P\left[(x-\eta)^2 + 4\lambda(x-\eta) + 1\right] + \lambda(r-1) + \eta \\
&= \frac{1}{4\lambda}\,\mathbb{E}_P\left[(x-\eta+2\lambda)^2\right] + \lambda(r-1) + \eta \\
&= \frac{1}{4\lambda}\,\mathbb{E}_P\left[(x-\tilde{\eta})^2\right] + \lambda(r+1) + \tilde{\eta}
\end{aligned}$$

Where $\tilde{\eta} := \eta - 2\lambda$, since η is unconstrained. Then, we can set the partial derivative with respect to λ to 0 to eliminate λ from $g(\lambda,\eta)$:

$$\begin{aligned}
\frac{\partial}{\partial\lambda} &= -\frac{1}{4\lambda^2}\,\mathbb{E}_P\left[(x-\tilde{\eta})^2\right] + (r+1) \\
0 &= -\frac{1}{4\lambda^2}\,\mathbb{E}_P\left[(x-\tilde{\eta})^2\right] + (r+1) \\
\lambda^2 &= \frac{1}{4(r+1)}\,\mathbb{E}_P\left[(x-\tilde{\eta})^2\right] \\
\lambda &= \frac{1}{2}\left(\frac{1}{r+1}\,\mathbb{E}_P\left[(x-\tilde{\eta})_+^2\right]\right)^{1/2}
\end{aligned}$$

where $(x-\tilde{\eta})_+ := \max(0, x-\tilde{\eta})$. For cleanliness, let $C := \mathbb{E}_P\left[(x-\tilde{\eta})_+^2\right]$. Plugging back into the original

definition we get:

$$\begin{aligned}\frac{1}{4\lambda}\,\mathbb{E}_P\left[(x-\tilde{\eta})^2\right]+\lambda(r+1)+\tilde{\eta} &= \frac{1}{2}\left(\frac{1}{r+1}C\right)^{-1/2}\cdot C+\frac{1}{2}\left(\frac{1}{r+1}C\right)^{1/2}(r+1)+\tilde{\eta}\\ &= \frac{1}{2}\left((r+1)C\right)^{1/2}+\frac{1}{2}\left((r+1)C\right)^{1/2}+\tilde{\eta}\\ &= \left(\frac{1}{r+1}C\right)^{1/2}+\tilde{\eta}\end{aligned}$$

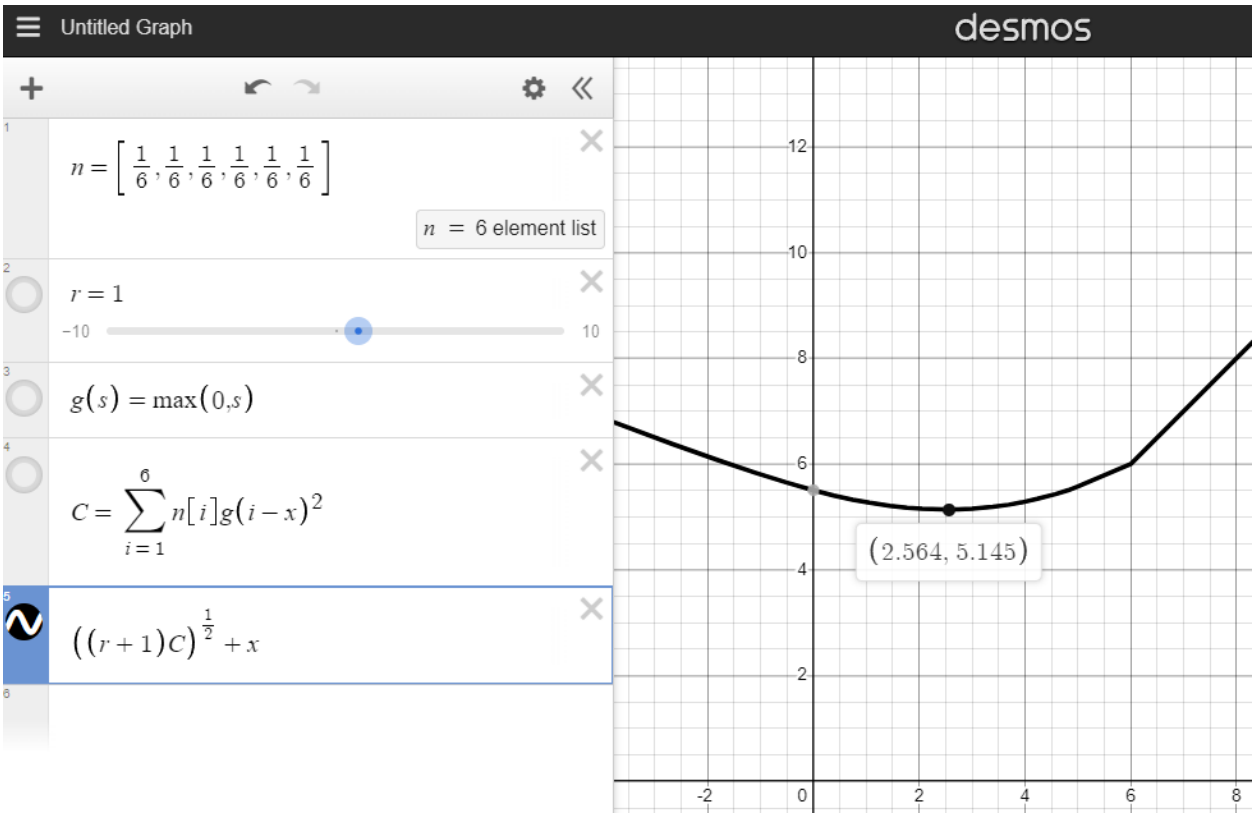
Summarizing these results, we have finally obtained:

$$\max_{Q}\left\{\mathbb{E}_Q[x] : D_f(Q\parallel P)\le r\right\}=\min_{\tilde{\eta}\in\mathbb{R}}\left\{\left(\frac{1}{r+1}\,\mathbb{E}_P\left[(x-\tilde{\eta})_+^2\right]\right)^{1/2}+\tilde{\eta}\right\} \tag{15}$$

The right-hand side of this equation is a single-variable convex function of $\tilde{\eta}$, for all $r>0$.

Interestingly, we can interpret the dual as meaning all sides of the die less than $\tilde{\eta}$ are given no weight, and the remaining sides are reweighted according to the squared term $(x-\tilde{\eta})^2$.

Example 2. *Take $r=1$ for example. Below is a plot of the right hand side from (15)*



The minimum is approximately 5.145, so this is approximately the largest value of $\mathbb{E}_Q[x]$ one could achieve by varying from the uniform distribution by at most $r=1$ in terms of the χ^2-divergence. In addition, note the elbow at $(6,6)$, which is in fact the maximum optimal $\tilde{\eta}$ and $\mathbb{E}_Q[x]$, when r is sufficiently large.

References

[1] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[2] Jonathan Borwein and Adrian Lewis. *Convex Analysis*. Springer, 2006.

[3] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach, 2016. URL: `https://arxiv.org/abs/1610.03425`.

[4] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. URL: `https://arxiv.org/abs/1806.08010`.