One possible motivation for a probability distribution π, is uniformity relative to a base measure, μ, which we can quantify with the Kullback-Leibler from μ to π,

$$\mathrm{KL}(\pi \mid\mid \mu) = \mathbb{E}_{\pi} \left[\log \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \right].$$

If both π and μ admit probability density functions relative to some reference measure, such as the counting measure on discrete spaces or the Lebesgue measure on a continuous space, then we can write this as

$$\mathrm{KL}(\pi \mid\mid \mu) = \int_X \mathrm{d}x \, \pi(x) \, \log \frac{\pi(x)}{\mu(x)}.$$

Identifying the most uniform probability distribution relative to the base measure then becomes a *variational* optimization problem of finding the measure that minimizes the Kullback-Leibler divergence subject to the measure integrating to one.

The solution to this variational optimization problem, however, many not be well-posed without additional constraints on π beyond normalization. From a probabilistic perspective any such constraint takes the form of an expectation of some function,

$$t_m = \mathbb{E}_{\pi}[T_m] = \int_X \mathrm{d}x \, \pi(x) \, T(x).$$

Any *smooth* variational solution subject to these constraints, if a solution exists at all, is given by the constrained Euler-Lagrange equations which in this case take the form

$$\frac{\partial L}{\partial \pi} = 0$$

where the *Lagrangian*, L, is given by

$$L = \pi(x) \, \log \frac{\pi(x)}{\mu(x)} + \sum_{m=1}^{M} \lambda_m \left(t_m - \int_X \mathrm{d}x \, \pi(x) \, T(x) \right) + \lambda_0 \left(1 - \int_X \mathrm{d}x \, \pi(x) \right).$$

The *Lagrange multipliers*, λ_m are implicitly defined by the $M + 1$ constraints and can be solved by imposing those constraints on any proposed solution.

Differentiating through the Lagrangian gives

$$\begin{aligned}
0 &= \frac{\partial L}{\partial \pi} \\
0 &= \log \frac{\pi(x)}{\mu(x)} + 1 - \sum_{m=1}^{M} \lambda_m \, T_m(x) - \lambda_0 \\
\log \frac{\pi(x)}{\mu(x)} &= \lambda_0 - 1 + \sum_{m=1}^{M} \lambda_m \, T_m(x) \\
\pi(x) &= \mu(x) \, \exp \left(\lambda_0 - 1 \right) \exp \left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right).
\end{aligned}$$

Imposing the normalization constraint gives

$$\begin{aligned}
1 &= \int_X \mathrm{d}x \, \pi(x) \\
&= \int_X \mathrm{d}x \, \mu(x) \exp\left(\lambda_0 - 1\right) \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right) \\
&= \exp\left(\lambda_0 - 1\right) \int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)
\end{aligned}$$

or

$$\exp\left(\lambda_0 - 1\right) = \frac{1}{\int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)} \equiv \frac{1}{Z(\lambda_1, \ldots, \lambda_M)}.$$

Substituting this into the variational solution finally gives

$$\pi(x) = \frac{\mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)}{\int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)} = \frac{\mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)}{Z(\lambda_1, \ldots, \lambda_M)}.$$

The Lagrange multipliers are implicit functions of the constraining expectation values and are identified by simultaneously solving the system of equations

$$\begin{aligned}
t_m &= \int_X \mathrm{d}x \, \pi(x) \, T(x) \\
&= \frac{\int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m'=1}^{M} \lambda_{m'} \, T_{m'}(x) \right) \, T_m(x)}{\int_X \mathrm{d}x \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_{m'} \, T_{m'}(x) \right)} \\
&= \frac{\frac{\partial}{\partial \lambda_m} \int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m'=1}^{M} \lambda_{m'} \, T_{m'}(x) \right)}{\int_X \mathrm{d}x \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_{m'} \, T_{m'}(x) \right)} \\
&= \frac{1}{Z(\lambda_1, \ldots, \lambda_M)} \frac{\partial}{\partial \lambda_m} Z(\lambda_1, \ldots, \lambda_M) \\
&= \frac{\partial}{\partial \lambda_m} \log Z(\lambda_1, \ldots, \lambda_M).
\end{aligned}$$

Such a solution, if it exists, defines the probability density function for a *maximum entropy* distribution relative to the given reference measure. Entropy here refers to the Kullback-Leibler divergence used to define the original variational objective function, although the reference is somewhat sloppy mathematically as entropy is defined on the *symplectic manifolds* that arise in classical mechanics and not a general space that we might be considering.

We can also avoid solving for the Lagrange multipliers and instead treat them as variables parameterizing an entire *family* of probability densities,

$$\pi(x; \lambda_1, \ldots, \lambda_M) = \frac{\mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)}{\int_X \mathrm{d}x \, \mu(x) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x) \right)}.$$

Any family of this form defines an *exponential family* relative to the reference measure with the *natural parameters*, $\{\lambda_1, \ldots, \lambda_M\}$. Note that depending on the choice of the $T_m(x)$ the natural parameters may be constrained to ensure self-consistent probability density functions.

Exponential families enjoy many convenient mathematical properties that make them particularly easy to manipulate in practice. For example the product of probability density functions within an exponential family falls within the same family,

$$\begin{aligned} \pi(x_1; \lambda_1, \ldots, \lambda_M) \times \pi(x_2; \lambda_1, \ldots, \lambda_M) &= \frac{\mu(x_1) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x_1) \right)}{Z(\lambda_1, \ldots, \lambda_M)} \frac{\mu(x_2) \, \exp\left(\sum_{m=1}^{M} \lambda_m \, T_m(x_2) \right)}{Z(\lambda_1, \ldots, \lambda_M)} \\ &= \frac{\mu(x_1) \, \mu(x_2) \, \exp\left(\sum_{m=1}^{M} \lambda_m (T_m(x_1) + T_m(x_2) \right)}{Z(\lambda_1, \ldots, \lambda_M)^2} \end{aligned}$$

Something something about how this is useful because we need to keep only $\sum_{n=1}^{N} T(x_n)$ instead of the entire data set $\{x_1, \ldots, x_N\}$.