The next step: what can we say generally about how \( p_n(t) \) approaches \( p_n \), especially for an isolated system.

Introduce a mathematical concept known as an \( f \)-divergence, which measures how similar are two prob. distributions.

Take two normalized distributions: \( \hat{p}', \hat{p} \) over \( n = 1, \ldots, \Omega \) states

\[
\sum_{n} p_n = \sum_{n} p'_n = 1 \quad p_n \geq 0 \quad p'_n \geq 0
\]
f-diverg. of $\hat{p}'$ from $\hat{p}$ is:

$$D_f(p' \| p) = \sum_n p_n f\left(\frac{p_n'}{p_n}\right)$$

where $f$ is any function which is
- convex, $f''(x) > 0$
- $f(1) = 0$

Properties of $D_f(p' \| p)$:

1) if $\hat{p}' = \hat{p} \Rightarrow D_f(p' \| p) = 0$ from $f(1) = 0$

2) if $\hat{p}' \neq \hat{p} \Rightarrow D_f(p' \| p) > 0$

Proof: First need Jensen's inequality:

$$\sum_n p_n f(x_n) \geq f\left(\sum_n p_n x_n\right)$$

for any prob. dist. $\{x_n, p_n\}$ and convex func. $f$

$\Theta = 2$:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

$x_1 \leq p_1 x_1 + p_2 x_2 \leq x_2$

$$= p_1 x_1 + (1-p_1) x_2$$

for any $p_1$

By induction, assume Jensen's for $\Theta = m$

show it works for $\Theta = m + 1$. 

\[ f \left( \sum_{n=1}^{m+1} p_n x_n \right) = f \left( p_1 x_1 + (1-p_1) \sum_{n=2}^{m+1} \frac{p_n}{1-p_1} x_n \right) \leq p_1 f(x_1) + (1-p_1) \sum_{n=2}^{m+1} \frac{p_n}{1-p_1} f(x_n) \]

Hence:

\[ D_f(p' \parallel p) = \sum_n p_n f \left( \frac{p_n'}{p_n} \right) \geq f \left( \sum_n p_n \frac{p_n'}{p_n} \right) = f(1) = 0 \]

Proof complete!
Remarkable theorem for any ergodic Markovian process (strongly connected network $\Rightarrow p_n^s$ exists, unique $p_n^s > 0$)

$\Rightarrow$ Monotonic f-divergence theorem (MFD)

$$
\frac{d}{dt} D_f (\hat{\mathbf{p}}(t) \parallel \hat{\mathbf{p}}^s) = 0
$$

with equality (zero slope) only when $\hat{\mathbf{p}}(t) = \hat{\mathbf{p}}^s$ at $t \to \infty$

$D_f (\hat{\mathbf{p}}(t) \parallel \hat{\mathbf{p}}^s)$ must monotonically approach zero as system approaches stationary state.

NOTE: works regardless of $\hat{\mathbf{p}}^s$ being an equilibrium state or NESS.

Proof:

$$
\frac{d}{dt} D_f (\hat{\mathbf{p}}(t) \parallel \hat{\mathbf{p}}^s)
$$

$$
= \frac{d}{dt} \sum_n p_n^s f \left( \frac{p_n(t)}{p_n^s} \right)
= \sum_n f' \left( \frac{p_n(t)}{p_n^s} \right) \frac{p_n(t)}{p_n^s} \Rightarrow p_n(t) = p_n^s x_n(t)
$$

$$
= \sum_{n,m} f'(x_n) \left[ W_{nm} p_m - W_{mn} p_n \right]
$$

using master equation $p_n'(t) = \sum_m \left[ W_{nm} p_m - W_{mn} p_n \right]$. 

\[
\sum_{n,m} f'(x_n) W_{nm} p_m - \sum_{n,m} f'(x_m) W_{nm} p_m
\]

we have interchanged labels in this sum.

\[
D_f (\hat{\psi}(t) | | \hat{\psi}^s) = \sum_{n,m} \frac{\rho'(x_n)}{\rho^s(x_m)} W_{nm} p_m^s [x_m f'(x_n) - x_m f'(x_m)]
\]

using \( p_m = x_m p_m^s \) \[ Eq. I \]

Now separately we will prove another identity needed for the proof. From the definition of the stationary state:

\[
0 = \sum_n [W_{nm} p_m^s - W_{mn} p_n^s]
\]

Let \( \Psi_n \) be an arbitrary set of numbers, \( n = 1, \ldots, N \).

\[
0 = \sum_n \psi_n \sum_m [W_{nm} p_m^s - W_{mn} p_n^s]
\]

\[
= \sum_{n,m} W_{nm} p_m^s \psi_n - \sum_{m,n} W_{nm} p_m^s \psi_m
\]

\[
0 = \sum_{n,m} W_{nm} p_m^s (\psi_n - \psi_m) \quad [Eq. II] \]

Choose \( \psi_n = f(x_n) - x_n f'(x_n) \)

and let us add Eq. II to Eq. I:

\[
\frac{d}{dt} D_f (\hat{\psi}(t) | | \hat{\psi}^s) = \sum_{n,m} W_{nm} p_m^s [f(x_m) - f(x_n)]
\]

\( \geq 0 \) by definition.
Final step: we need to prove the term in brackets \([\_\_\_]\) is always negative for convex function \(f\)

\[ f(x_m) \geq f(x_n) + (x_m - x_n)f'(x_n) \]

with equality only when \(x_m = x_n\)

To show this, start with Jensen's inequality with \(\Theta = 2\):

\[
(1-\varepsilon)f(x_n) + \varepsilon f(x_m) \\
\geq f\left((1-\varepsilon)x_n + \varepsilon x_m\right) \quad \text{[equal only if } x_m = x_n\text{]} \\
= f(x_n + \varepsilon(x_m - x_n)) \\
\Rightarrow \varepsilon f(x_m) \geq \varepsilon f(x_n) + f(x_n + \varepsilon(x_m - x_n)) - f(x_n)
\]

Divide by \(\varepsilon\) and take limit as \(\varepsilon \to 0\):

\[ f(x_m) \geq f(x_n) + f'(x_n)(x_m - x_n) \]

\(\cdot\) Thus we have shown that \(\frac{d}{dt}D_f(\overset{^t}{\mathbf{p}}(t) \| \mathbf{p}_s) \leq 0\)

Equality can only occur if \(x_n = x_m\) for all pairs of states \((n, m)\) where \(W_{nm} \neq 0\). But in a strongly connected network this implies all \(x_n\) are equal to each other.

Since \(x_n = \frac{P_n(t)}{P_n^s}\) the only way this can occur is if \(p_n(t) = P_n^s\) for all \(n\), \([\text{by normalize}]\)

Hence \(\frac{d}{dt}D_f(\overset{^t}{\mathbf{p}}(t) \| \mathbf{p}_s) = 0\) when \(t \to \infty\).
One particular choice of $f$ is very useful:

$$f(x) = x \ln x$$

$$\Rightarrow D_{KL}(p \parallel p^s) = \sum_n p^*_n \ln \left( \frac{p^*_n}{p_n^s} \right) \equiv \text{Kullback-Leibler divergence of } p \text{ from } p^s$$

Why is this physically attractive? [in information theory: how much info is lost when $p^s$ is used to approx. $p$]

If we have two separate, isolated systems, both approaching equilibrium:

$A$ and $B$

$C = A + B$

$P^n_A(t) \rightarrow P^{As}_n$

$P^n_B(t) \rightarrow P^{Bs}_n$

$P^{C}_{n,m}(t) = P^{A}_{n}(t) P^{B}_{m}(t) \rightarrow P^{As}_{n}P^{Bs}_{m} \equiv P^{Cs}_{n,m}$

Intuitively: $(\text{distance of } A \text{ from equil.)} + (\text{distance of } B \text{ from equil.)}) = (\text{dist. of } C \text{ from equil.)}$

$$D_{KL}(p^c \parallel p^{Cs}) = \sum_{n,m} p^n_A p^m_B \ln \left( \frac{p^n_A p^m_B}{p^{As}_n p^{Bs}_m} \right)$$

$$= \sum_{n,m} p^n_A p^m_B \ln \left( \frac{p^n_A}{p^{As}_n} \right) + \sum_{n,m} p^n_A p^m_B \ln \left( \frac{p^m_B}{p^{Bs}_m} \right)$$

$$= \sum_{n} p^n_A \ln \left( \frac{p^n_A}{p^{As}_n} \right) + \sum_{m} p^m_B \ln \left( \frac{p^m_B}{p^{Bs}_m} \right)$$

$$= D_{KL}(p^A \parallel p^{As}) + D_{KL}(p^B \parallel p^{Bs})$$
Is this the only unique $f(x)$ with this property?

No, for example $f(x) = -\ln x$ gives $D_{KL}(p^x \| p)$

or $f(x) = \alpha \times \ln x \quad -(1-\alpha) \times \ln x \quad 0 \leq \alpha \leq 1$

also works.

More exotic examples [These are the only choices.]