This document provides a line-by-line, fully expanded derivation of how the Free Energy Principle (FEP) emerges from fundamental physical laws (conservation of probability) and thermodynamic constraints (stochastic thermodynamics, local detailed balance), together with biologically relevant mechanisms (gating, memory, hierarchy). The derivation makes explicit every step involved, culminating in the well-known variational characterization of nonequilibrium steady states. Biological systems often operate across multiple scales of organization, where processes at a lower level (e.g., molecular) influence and are influenced by higher levels (e.g., cellular or organismal). This hierarchical organization introduces feedback loops that span scales, coupling local dynamics with global constraints. For example, in visual processing, neural layers iteratively refine sensory inputs through top-down and bottom-up signals. To account for such multi-scale interactions, we incorporate hierarchical feedback terms into the flux decomposition and continuity equations, as detailed in Section 4.1.

1 Overview

We aim to show how:

1. *Probability conservation* (continuity equation) governs the evolution of a probability density $p(x, t)$.

2. *Stochastic thermodynamics* introduces an entropy production rate and imposes local detailed balance on transition rates, ensuring thermodynamic consistency.

3. *Biological constraints* such as gating (context-dependent transitions) and memory (non-Markovian dynamics) modify the probability flux.

4. *A free energy functional* $F[p]$ naturally emerges, whose minimization in steady state is equivalent to setting flux divergences to zero (homeostasis).

We follow a step-by-step approach, making explicit the algebra at each stage.

2 Conservation of Probability

2.1 Continuity Equation

Consider a system with states labeled by $x \in \Omega$, where Ω is the state space (which can be continuous for many biological systems). Let $p(x, t)$ denote the probability density of the system being in state x at time t. Probability conservation (no net loss or gain of probability) implies that the rate of change of $p(x, t)$ is governed by the continuity equation:

$$\frac{\partial p(x,t)}{\partial t} + \nabla \cdot J(x,t) = 0. \tag{1}$$

Here,

- $J(x,t)$ is the probability flux (a vector field on Ω, if Ω is spatial or analogous if Ω is more abstract).
- The symbol ∇ denotes the gradient with respect to x, and $\nabla \cdot (\,\cdot\,)$ the corresponding divergence.

The flux $J(x,t)$ reflects both diffusive (random) and drift (systematic) components of motion. In general, any biologically or physically motivated constraints can be built directly into the form of $J(x,t)$.

3 Stochastic Thermodynamics and Entropy Production

3.1 Definition of Entropy Production

Biological systems typically operate far from thermodynamic equilibrium and constantly exchange energy and matter with their environment. In stochastic thermodynamics, the *entropy production rate* $\dot{\Sigma}(t)$ is a crucial measure of irreversibility:

$$\dot{\Sigma}(t) = \int_{\Omega} J(x,t) \cdot \nabla \ln\Big(\tfrac{p(x,t)}{m(x)}\Big)\, dx, \tag{2}$$

where

- $p(x,t)$ is the instantaneous probability density of the system.
- $m(x)$ is a reference (or stationary) distribution. In many contexts (e.g. near equilibrium), $m(x)$ could be a Boltzmann (Gibbs) distribution $\propto e^{-U(x)/(k_B T)}$.
- The integrand $J(x,t) \cdot \nabla \ln\big(\tfrac{p(x,t)}{m(x)}\big)$ is the local contribution to entropy production.

Note that $\dot{\Sigma}(t)$ represents the total entropy production rate, which is always non-negative in physically consistent systems.

3.2 Local Detailed Balance

The condition of *local detailed balance* encapsulates the microscopic reversibility required by thermodynamics. If $W(x \to x')$ is the transition rate from state x to x', then local detailed balance states:

$$\frac{W(x \to x')}{W(x' \to x)} = \exp\Big(-\frac{\Delta F(x,x')}{k_B T}\Big), \tag{3}$$

where

- $\Delta F(x,x')$ is the free-energy difference associated with the transition $x \to x'$.
- k_B is the Boltzmann constant.

- T is the temperature.

This requirement ensures that in equilibrium (i.e. when $p(x,t)$ is the stationary Boltzmann-type distribution $m(x)$), detailed balance holds and there is no net flux in the system. Far from equilibrium, one can still track how much the system deviates from equilibrium via $\dot{\Sigma}(t)$.

4 Flux Decomposition with Gating and Memory

Biological systems often involve additional structure beyond mere Markovian transitions:

1. **Gating Functions:** $g(x)$ can represent threshold-dependent or context-dependent restrictions on transitions (e.g. neural spike thresholds, enzymatic gating).
2. **Memory Effects:** Non-Markovian or time-delayed feedbacks arise in processes such as gene regulatory networks or other feedback loops. A memory kernel $K(\tau; x', x)$ captures how past states (x' at time $t-\tau$) influence the current flux at state x.

A general probability flux $J(x,t)$ that encodes such biological constraints can be written as:

$$J(x,t) = -\,D\,g(x)\,p(x,t)\,\nabla\mu(x,t) \;+\; \int_0^\infty\!\!\int_\Omega K(\tau; x', x)\,p(x', t-\tau)\,dx'\,d\tau, \tag{4}$$

where

- D is a diffusion coefficient or mobility factor.
- $g(x)$ is a gating function enforcing that the flux vanishes (or is severely reduced) for certain values of x.
- $\mu(x,t)$ is a *chemical potential*, which we will show is related to the functional derivative of the system's free energy.
- The second term captures the integral of all delayed influences up to time $t-\tau$ through the kernel $K(\tau; x', x)$.

4.1 Hierarchical Feedback and Multi-Scale Organization

To model hierarchical feedback, we introduce an additional coupling term $H(x, x'; \ell)$, where ℓ represents a hierarchical layer index. This term allows the flux at one scale to depend on the states of other layers, reflecting top-down and bottom-up interactions. The total flux $J(x,t)$ is extended as:

$$J(x,t) = -Dg(x)p(x,t)\nabla\mu(x,t) + \int_0^\infty \int_\Omega K(\tau; x', x)p(x', t-\tau)\,dx'\,d\tau + \sum_\ell H(x, x'; \ell), \tag{5}$$

where $H(x, x'; \ell)$ is defined as:

$$H(x, x'; \ell) = \alpha_\ell \nabla_x \left(f_\ell(x) p_\ell(x', t)\right), \tag{6}$$

with $f_\ell(x)$ representing the feedback function for layer ℓ and α_ℓ a coupling coefficient that determines the strength of the interaction. The feedback term can be symmetric (reciprocal influence between layers) or asymmetric (dominance of one scale), depending on the specific biological context.

4.2 Definition of Chemical Potential

We define the chemical potential $\mu(x,t)$ as the functional derivative of a *free energy functional* $F[p]$ with respect to the probability density $p(x,t)$:

$$\mu(x,t) = \frac{\delta F[p]}{\delta p(x,t)}. \tag{7}$$

The next sections make explicit why $F[p]$ can be regarded as a "free energy" and how its form emerges from thermodynamic and probabilistic considerations.

5 Variational Principle and the SPDE

5.1 Substituting the Flux into the Continuity Equation

Recalling the continuity equation (1) (reproduced here for clarity):

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot J(x,t), \tag{8}$$

and substituting our expression for the flux $J(x,t)$ from (4), we obtain:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot \Big[-D\, g(x)\, p(x,t)\, \nabla \mu(x,t) \Big] \; - \; \nabla \cdot \Big[\int_0^\infty \!\! \int_\Omega K(\tau; x', x)\, p(x', t-\tau)\, dx'\, d\tau \Big].$$

To make this more explicit, we note that

$$-\nabla \cdot \Big[-Dg(x) p(x,t) \nabla \mu(x,t) \Big] = \nabla \cdot \Big[Dg(x) p(x,t) \nabla \mu(x,t) \Big].$$

For the memory term, we recognize that probability conservation under memory may require an additional corrective term (often denoted $\Phi_{\text{mem}}(x,t)$) to ensure that all probability fluxes balance properly over time (especially if $K(\tau; x', x)$ does not integrate to zero). Hence, in the most general form, one writes:

$$\begin{aligned} \frac{\partial p(x,t)}{\partial t} = \nabla \cdot \Big[D\, g(x)\, p(x,t)\, \nabla \mu(x,t) \Big] \; + \; \int_0^\infty \!\! \int_\Omega K(\tau; x', x)\, p(x', t-\tau)\, dx'\, d\tau \\ - \; \Phi_{\text{mem}}(x,t), \end{aligned} \tag{9}$$

where the term $\Phi_{\text{mem}}(x,t)$ is determined self-consistently to enforce probability conservation (i.e. so that the integral of $\frac{\partial p}{\partial t}$ over Ω vanishes).

Equation (9) is a *stochastic partial differential equation* (SPDE) once one accounts for any noise sources or stochastic forcing; it incorporates gating, memory, and thermodynamic consistency.

Incorporating hierarchical feedback modifies the continuity equation to include cross-scale interactions. Substituting the extended flux $J(x,t)$ into the continuity equation gives:

$$\frac{\partial p(x,t)}{\partial t} = \nabla\cdot[Dg(x)p(x,t)\nabla\mu(x,t)] + \int_0^\infty \int_\Omega K(\tau;x',x)p(x',t-\tau)\,dx'\,d\tau + \sum_\ell \nabla\cdot H(x,x';\ell) - \Phi_{\text{mem}}(x,t). \tag{10}$$

This term $\nabla \cdot H(x,x';\ell)$ encodes the influence of hierarchical feedback, allowing the variational principle to capture emergent behaviors arising from multi-scale dynamics. The presence of $H(x,x';\ell)$ introduces an additional layer of complexity in minimizing the free energy functional $F[p]$, reflecting the interplay between scales.

6 Emergence of the Free Energy Functional

6.1 Form of the Free Energy Functional

We now show how the functional $F[p]$ arises naturally from considerations of:

1. **Relative entropy (KL divergence):** comparing $p(x)$ to a reference distribution $m(x)$.
2. **System-specific potential energy:** $U(x)$.

Concretely, we define:

$$F[p] = \int_\Omega p(x)\,\ln\!\Big(\tfrac{p(x)}{m(x)}\Big)\,dx \;+\; \int_\Omega p(x)\,U(x)\,dx. \tag{11}$$

Breaking this down:

- The term $\displaystyle\int p(x)\ln\big(p(x)/m(x)\big)\,dx$ measures the (Kullback–Leibler) divergence of $p(x)$ from $m(x)$. If $m(x)$ is taken to be the Boltzmann distribution $m(x) \propto e^{-U(x)/(k_BT)}$, this term represents an *entropy* difference between p and m.
- The term $\displaystyle\int p(x)\,U(x)\,dx$ corresponds to the *internal energy* (or potential energy) of the system under constraints $U(x)$ (e.g. external fields, chemical potentials).

Thus, $F[p]$ is a natural *nonequilibrium free energy*, combining entropy and internal energy contributions.

6.2 Chemical Potential via Functional Derivative

To see that $\mu(x,t) = \frac{\delta F[p]}{\delta p(x,t)}$ is precisely the "chemical potential" in (7), we compute the functional derivative of (11):

$$F[p] = \int_\Omega p(x)\, \ln p(x)\, dx \;-\; \int_\Omega p(x)\, \ln m(x)\, dx \;+\; \int_\Omega p(x)\, U(x)\, dx.$$

Vary $F[p]$ with respect to $p(y)$:

$$\frac{\delta}{\delta p(y)}\Big[\int_\Omega p(x) \ln p(x)\, dx\Big] = \ln p(y) + 1,$$

$$\frac{\delta}{\delta p(y)}\Big[-\int_\Omega p(x) \ln m(x)\, dx\Big] = -\ln m(y),$$

$$\frac{\delta}{\delta p(y)}\Big[\int_\Omega p(x) U(x)\, dx\Big] = U(y).$$

Summing these contributions:

$$\frac{\delta F[p]}{\delta p(y)} = \big(\ln p(y) + 1\big) \;-\; \ln m(y) \;+\; U(y).$$

Since an additive constant (such as $+1$) in the chemical potential has no effect on its gradient or on the ensuing flux dynamics, we usually absorb it into an overall reference level. Hence, one often writes:

$$\mu(y) = \ln\Big(\tfrac{p(y)}{m(y)}\Big) \;+\; U(y). \tag{12}$$

This identification confirms that the gradient of μ in Eq. (4) depends on the gradient of $\ln(p/m)$ plus the gradient of U. Therefore, the flux $J(x,t)$ in (4) enforces motion in the direction that decreases $F[p]$, all else being equal.

7 Connection to the Standard Free Energy Principle (FEP)

7.1 Steady-State Conditions and the Minimization of $F[p]$

In nonequilibrium steady state, the probability distribution $p^*(x)$ satisfies

$$\frac{\partial p^*(x)}{\partial t} = 0.$$

From the continuity equation (1) (or the SPDE (9)), this implies

$$\nabla \cdot J^*(x) = 0, \tag{13}$$

where $J^*(x)$ is the steady-state flux.

- In many physically relevant scenarios (particularly when detailed balance or near-detailed-balance conditions hold), the steady-state flux itself can vanish: $J^*(x) = 0$.
- More generally, in a nonequilibrium steady state, one can still often characterize $p^*(x)$ as a minimizer (or critical point) of the free energy functional, subject to constraints imposed by nonzero currents.

Nonetheless, a common simplification in the standard FEP reading is to focus on a near-equilibrium or effectively equilibrium-like regime, where $J^*(x) = 0$ exactly, giving:

$$J^*(x) = -\,D\,g(x)\,p^*(x)\,\nabla\mu^*(x) \;=\; 0. \tag{14}$$

Hence, $p^*(x)$ satisfies

$$\nabla\mu^*(x) = \nabla\Big(\ln\big(p^*(x)/m(x)\big) + U(x)\Big) = 0,$$

yielding

$$p^*(x) \;\propto\; m(x)\,\exp\big(-U(x)\big),$$

or the more familiar Boltzmann-like form if $m(x)$ itself is $\propto e^{-U(x)/(k_BT)}$.

7.2 Biological Interpretation: Minimizing Variational Free Energy

Within the *Free Energy Principle* framework in theoretical biology and neuroscience, one interprets $p^*(x)$ as the organism's (or system's) probabilistic representation of hidden states of the world. The principle states that the organism *self-organizes* so as to minimize its free energy $F[p]$, thereby resisting disorder and maintaining homeostasis. We see from the above derivation that:

$$(\text{Minimizing } F[p]) \iff \nabla \cdot J^*(x) = 0,$$

with $J^*(x)$ given by physically consistent fluxes (including gating, memory, etc.). Thus, the FEP emerges as a *variational principle* from the foundational laws of stochastic thermodynamics and probability conservation, rather than being introduced as a mere *ad hoc* postulate.

8 Conclusion

We have expanded, step by step, how a *free energy functional* $F[p]$ arises from fundamental principles (probability conservation and thermodynamics) and how its minimization at steady state connects directly to the standard Free Energy Principle. Crucially,

1. We began with the continuity (probability conservation) equation.
2. We incorporated local detailed balance to ensure thermodynamic consistency.
3. We allowed for biologically relevant modifications to the flux (gating, memory, hierarchy).

4. The free energy functional $F[p]$ then follows naturally, with its functional derivative defining a chemical potential μ.

5. The resulting SPDE and steady-state conditions show how the system organizes around minima of $F[p]$, consistent with the FEP.

Hence, one obtains a rigorous, biologically grounded explanation of how systems maintain order (i.e. homeostasis) under nonequilibrium conditions. In this sense, the FEP is not an independent assumption but rather a corollary of more fundamental dynamical and thermodynamic laws. The inclusion of hierarchical feedback provides a pathway to bridge the dynamics across scales, integrating local transitions with global constraints. For example, in neural systems, this allows the modeling of iterative refinements between sensory inputs and perceptual interpretations. Future work could explore specific applications of hierarchical feedback, such as how multi-scale interactions stabilize homeostasis or generate emergent behaviors in complex systems.