8.333 Fall 2025 Recitation 3

Jessica Metzger
jessmetz@mit.edu | Office hours: Tuesday 4-5pm (2-132)

These notes are largely a conglomeration of the previous years' recitation notes by Julien Tailleur, Amer Al-Hiyasat, and Sara Dal Cengio.

References. All the essential information in these recitations can be found in Chapter 2 of Mehran Kardar's *Statistical Physics of Particles.*

This recitation will comprise assorted topics relevant to the lectures and psets.

I. PRINCIPLE OF MINIMAL INFORMATION: THE WALLIS DERIVATION

Recall that in lecture we derived a principle of minimal information to find the "least biased" distribution. We found a quantity called the "surprise" of a sample that is additive for independent samples and is larger for more rare samples. For a sample with probability p, we found that the surprise must be proportional to $-\ln p$. The conclusion was that the least-biased probability distribution p_n over states n should minimize the average surprise:

$$S = -k_B \sum_n p_n \ln p_n \ . \tag{1}$$

There is another way to derive this principle, due to Graham Wallis and E. T. Jaynes, which I'll go over here. Imagine a group of N equally-sized buckets. Blindfolded, you throw W balls into these buckets. Each throw is independent of the others. Overall, bucket i receives n_i balls. You use this to construct an empirical probability distribution, letting $p_i \equiv n_i/W$ be the probability of box i.

However, suppose you are only interested in probability distributions that satisfy some constraint. For example, if the buckets have different "energies", suppose you only care about probability distributions whose average energy $\langle E \rangle = \sum_i p_i E_i$ is within some interval $[E, E + \Delta E]$. To enforce a constraint, after you do your sampling, you throw away the empirical probability distribution if it doesn't satisfy the constraint. You keep sampling, and throwing out inadmissible samples, until you have many samples, yielding many independent empirical probability distributions that all satisfy the constraint.

Note that there are multiple ways to generate the same probability distribution $\{p_i\}$, or occupancies $\{n_i\}$: the balls could land in the same boxes but in a different order. The number of different ways to generate a sample is given by its degeneracy Ω, which is given by the multinomial coefficient:

$$\Omega[\{p_i\}] = \frac{W!}{n_1! n_2! \dots n_N!} = \frac{W!}{(p_1 W)!(p_2 W)! \dots (p_N W)!} \ . \tag{2}$$

We can now make the crucial observation that samples with higher degeneracy are more likely to be observed: the most probable probability distribution is the one that maximizes Ω. Another way to maximize Ω is to maximize the quantity

$$S[\{p_i\}] \equiv \frac{1}{W} \ln \Omega[\{p_i\}] = \frac{1}{W} \ln \left(\frac{W!}{(p_1 W)!(p_2 W)! \dots (p_N W)!} \right) = \frac{1}{W} \left(\ln W! - \sum_{i=1}^{N} \ln(p_i W)! \right) . \tag{3}$$

To generate an accurate probability distribution, you need to make very large samples, so $W \to \infty$. But in this limit, S can be simplified using **Stirling's approximation**:

Lemma I.1: Stirling's approximation

The logarithm of the factorial can be approximated as

$$\ln W! = W \ln W - W + \mathcal{O}(\ln W) . \tag{4}$$

Using this approximation, we find

$$S[\{p_i\}] \approx \frac{1}{W} \left(W \ln W - W - \sum_{i=1}^{N} \Big[p_i W \ln(p_i W) - p_i W \Big] \right) = \ln W - 1 - \sum_{i=1}^{N} \Big[p_i \ln(p_i W) - p_i \Big] \tag{5}$$

$$= \ln W - 1 - \sum_{i=1}^{N} \Big[p_i \ln p_i \Big] - \sum_{i=1}^{N} \Big[p_i \Big] \ln W - \sum_{i=1}^{N} \Big[- p_i \Big] = \ln W - 1 - \sum_{i=1}^{N} \Big[p_i \ln p_i \Big] - \ln W + 1 \tag{6}$$

$$= - \sum_{i=1}^{N} p_i \ln p_i . \tag{7}$$

This is simply the Shannon entropy of the system. Thus, because we are most likely to observe the highest-degeneracy probability distribution, and because the highest-degeneracy probability distribution also has the highest entropy, we are most likely to observe the maximum-entropy probability distribution. Of course, all of this is restricting to the set of distributions that satisfy the constraint.

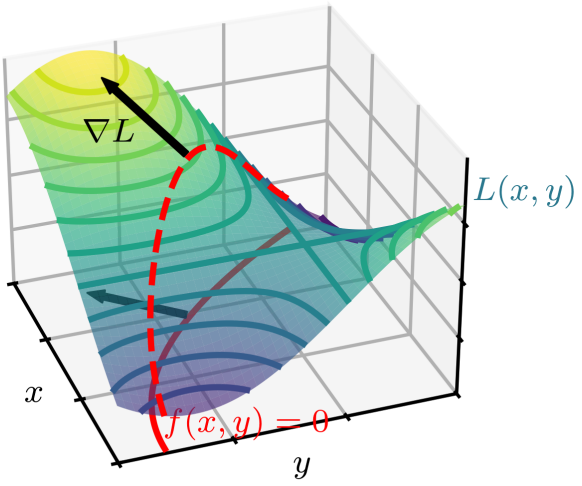
II. LAGRANGE MULTIPLIERS

A. One constraing

How do we maximize a multidimensional function while satisfying a constraint? We use the method of Lagrange multipliers.

Consider a toy example in 2d. We would like to maximize the function $L(x, y)$ with respect to the variables x and y, but we are forced to adhere to some constraint; for instance, $f(x, y)$ must equal zero for some f. In other words, we have to find the point along the line $f(x, y) = 0$ such that $L(x, y)$ is maximum.

At the maximum of L along the constraint, the constraint curve must be parallel to the contour line of L. But the constraint curve is simply a contour line of f. In other words, the gradient of the constraint function f must be parallel to the gradient of L. This is illustrated in the following figure:



If the gradients of L and f are parallel, then we can write $\nabla L(x,y) = \lambda \nabla f(x,y)$ for some scalar λ. This justifies defining a new lagrangian

$$\mathcal{L}(x,y,\lambda) \equiv L(x,y) - \lambda f(x,y) \tag{8}$$

and maximizing it with respect to x, y, and λ. We see that, indeed,

$$\nabla_{x,y}\mathcal{L} = 0 \qquad \Longrightarrow \qquad \nabla_{x,y} L = \lambda \nabla_{x,y} f \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \qquad \Longrightarrow \qquad f(x,y) = 0\,. \tag{10}$$

B. Multiple constraints

What if there are multipe constraints? As an example, consider the problem of maximizing a higher-dimensional function $L(\vec{x})$ with respect to $\vec{x}$ while also requiring that $f_1(\vec{x}) = 0$ and $f_1(\vec{x}) = 0$. We would define the new function

$$\mathcal{L}(\vec{x};\lambda_1,\lambda_2) \equiv L(\vec{x}) - \lambda_1 f_1(\vec{x}) - \lambda_2 f_2(\vec{x})\,. \tag{11}$$

Then, extremizing $\mathcal{L}$ with respect to $\vec{x}$, λ_1, and λ_2 results in all constraints being satisfied, along with

$$\nabla L = \lambda_1 \nabla f_1 + \lambda_2 \nabla f_2\,. \tag{12}$$

Or, the gradient of L is a linear combination of the gradients of f_1 and f_2. The vector space spanned by the gradients of the constraints is, locally, the "forbidden" space in which the constraints prevent us from moving. Thus, if ∇L lives in this space, the "allowed" space is perpendicular to ∇L. In other words, if we move along the constraint, L doesn't change, and thus at this point it is extremized.

III. FUNCTIONAL DERIVATIVES

We have considered optimization problems of finite or countable dimension, e.g. the N-dimensional problem of maximizing $S[\{p_i\}] = -\sum_{i=1}^{N} p_i \ln p_i$ with respect to each p_i. But what if the dimension of the space is infinite and uncountable; for instance, how do we maximize something over the space of real functions?

For this, we use the functional calculus, a functional generalization of vector calculus. In functional calculus, functions are replaced with functionals, the variables are replaced with functions, and indices are replaced with variables. The analogy between vector calculus and functional calculus is summarized in the following table:

Vector calculus	Functional calculus
function $f(\vec{x})$	functional $\mathcal{F}\big[f(\vec{x})\big]$
variable $\vec{x}$	function $f(\vec{x})$
index i	variable $\vec{x}$
partial derivative $\partial f/\partial x_i$	functional derivative $\delta\mathcal{F}/\delta f(\vec{x})$

As with the partial derivative in vector calculus, the functional derivative is constructed by considering the variation of a functional when only one "coordinate" is varied; i.e. we vary f at only one position x. For a functional $\mathcal{F}[f]$, we define

$$\frac{\delta\mathcal{F}\big[f(\vec{y})\big]}{\delta f(\vec{x})} \equiv \lim_{\varepsilon\to 0}\frac{\mathcal{F}\big[f(\vec{y})+\varepsilon\delta(\vec{y}-\vec{x})\big]-\mathcal{F}\big[f(\vec{y})\big]}{\varepsilon}\ . \tag{13}$$

(This isn't always well-defined because of the pathological properties of the delta function, and in math a different construction using test functions is used. However, for our purposes, this will always work.)

We can derive some important relations using the definition (13). For example, if the functional can be written as the integral of a function $L(x,f,f')$, i.e. as

$$\mathcal{F}[f(x)] = \int dx L\big(x,f(x),f'(x)\big)\ , \tag{14}$$

then we can calculate

$$\frac{\delta\mathcal{F}[f(y)]}{\delta f(x)} = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\bigg[\int dy\Big(L\big(y,f(y)+\varepsilon\delta(y-x),f'(y)+\varepsilon\delta'(y-x)\big)\ -\ L\big(y,f(y),f'(y)\big)\Big)\bigg] \tag{15}$$

$$= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\bigg[\int dy\Big(L(y,f(y),f'(y))+\varepsilon\frac{\partial L}{\partial f}(y)\delta(y-x)+\varepsilon\frac{\partial L}{\partial f'}(y)\delta'(y-x)+\mathcal{O}(\varepsilon^2)-L(y,f(y),f'(y))\Big)\bigg] \tag{16}$$

$$= \int dy\bigg[\delta(y-x)\frac{\partial L}{\partial f}(y)+\delta'(y-x)\frac{\partial L}{\partial f'}(y)\bigg] \tag{17}$$

$$= \int dy\bigg[\delta(y-x)\frac{\partial L}{\partial f}(y)-\delta(y-x)\frac{\partial}{\partial y}\Big(\frac{\partial L}{\partial f'}(y)\Big)\bigg] \tag{18}$$

$$= \frac{\partial L}{\partial f}(x)-\frac{\partial}{\partial x}\Big(\frac{\partial L}{\partial f'}(x)\Big)\ . \tag{19}$$

This leads us to the Euler-Lagrange equation of classical mechanics. (The chain rule given as a hint in the pset follows from this.)

We can also take the functional derivative of a *function* f with respect to itself. Making f look more like a functional by writing

$$f(y) = \int dz f(z)\delta(y-z)\ , \tag{20}$$

we can use Eq. (19) with $L(x,f(x)) = f(x)\delta(y-x)$ (forgetting that y is a variable) to find

$$\frac{\delta f(y)}{\delta f(x)} = \frac{\partial L}{\partial f}(x) = \delta(y-x)\ . \tag{21}$$

IV. LIOUVILLE'S THEOREM

We will now prove Liouville's theorem, first for single-particle dynamics then multi-particle dynamics, only in 1d. (The generalization to 3d will follow straightforwardly.)

A. Single-particle

First consider a single particle moving in 1-dimensional space with coordinate q, momentum p, and Hamiltonian $H(q,p)$. Suppose it starts at an initial phase space coordinates (q,p), and consider its evolution over a small interval of time Δt.

Hamilton's equations of motion dictate that the phase space coordinates evolve as

$$q \quad \to \quad q' = q + \dot{q}\Delta t + \mathcal{O}(\Delta t^2) = q + \frac{\partial H}{\partial p}\Delta t + \mathcal{O}(\Delta t^2) \tag{22}$$

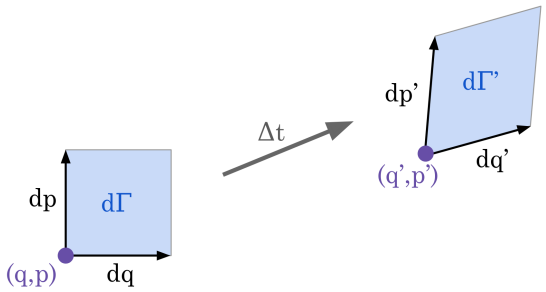
$$p \quad \to \quad q' = p + \dot{p}\Delta t + \mathcal{O}(\Delta t^2) = p - \frac{\partial H}{\partial q}\Delta t + \mathcal{O}(\Delta t^2)\ . \tag{23}$$

Now consider a small volume of phase space, $d\Gamma = [q, q+dq] \times [p, p+dp]$, located at (q,p). Its volume is $|d\Gamma| = dqdp$. Under the Hamiltonian evolution, it moves to a new volume $d\Gamma'$. It may shift, rotate, and stretch. But does its volume change?

The volume of the $d\Gamma'$ can be related to $|d\Gamma|$ using the Jacobian determinant of the infinitesimal transformation (22)-(23). That is,

$$|d\Gamma'| = (\det \mathcal{J})|d\Gamma|\ , \qquad \text{where } \mathcal{J} = \begin{pmatrix} \partial q'/\partial q & \partial q'/\partial p \\ \partial p'/\partial q & \partial p'/\partial p \end{pmatrix}\ . \tag{24}$$

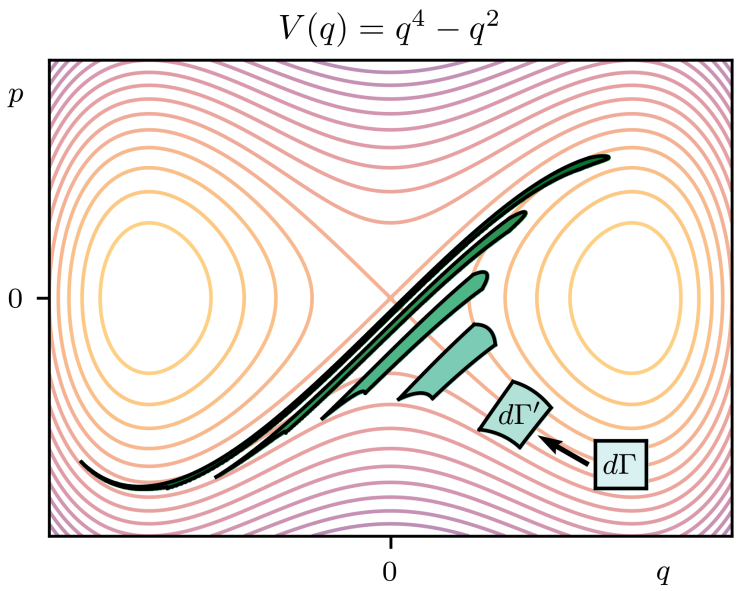
See the following schematic:



The Jacobian matrix is found using the infinitesimal transformation (22)-(23) to be

$$\det \mathcal{J} = \det \begin{pmatrix} 1 + \Delta t \partial^2 H/\partial p \partial q & \Delta t \partial^2 H/\partial p^2 \\ -\Delta t \partial^2 H/\partial q^2 & 1 - \Delta t \partial^2 H/\partial q \partial p \end{pmatrix} = 1 + \mathcal{O}(\Delta t^2) \quad \Longrightarrow \quad |d\Gamma'| = |d\Gamma|. \tag{25}$$

Thus the volume of $d\Gamma$ is unchanged as it evolves under the Hamiltonian dynamics. This evolution is visualized in the following figure:



1. Evolution of the probability density

One consequence of this is the conservation of probability density along a trajectory. Consider an ensemble of independently-evolving single particle systems. Suppose the probability density of configurations in the ensemble is given by $\rho(q,p;t)$. (This is a probability density over the coordinates (q,p), parametrized by the time t. At each time t, there is a new probability density $\rho(q,p;t)$.) So at time t, there is a probability $\sim \rho(q,p;t)dqdp$ of being in the box $[q,q+dq]\times[p,p+dp]$.

Think from the perspective of a single particle. In a small neighborhood of configuration space near you, how many other particles from the ensemble do you expect there to be? This is proportional to $\sim \rho(q,p;t)dqdp$. What about at time $t+\Delta t$? You move to $q' = q + \dot{q}\Delta t$ and $p' = p + \dot{p}\Delta t$. The neighborhood around you also shifts by the same amount, and deforms according to the Jacobian matrix $\mathcal{J}$ given in Eq. (24). Thus the number of neighbors transforms like

$$\begin{aligned}
\#\text{ Neighbors} &\propto \rho(q,p;t)|d\Gamma| && (26)\\
&\xrightarrow[\Delta t]{} \rho(q+\dot{q}\Delta t, p+\dot{p}\Delta t; t+\Delta t)|d\Gamma'| && (27)\\
&= \left[\rho(q,p;t) + \Delta t\dot{q}\frac{\partial\rho}{\partial q} + \Delta t\dot{p}\frac{\partial\rho}{\partial p} + \Delta t\frac{\partial\rho}{\partial t}\right]|d\Gamma| && (28)\\
&= \left[\rho(q,p;t) + \Delta t\left(\frac{\partial\rho}{\partial q}\frac{\partial H}{\partial p} - \frac{\partial\rho}{\partial p}\frac{\partial H}{\partial q} + \frac{\partial\rho}{\partial t}\right)\right]|d\Gamma| && (29)\\
&= \Big[\rho(q,p;t) + \{\rho,H\} + \partial_t\rho\Big]|d\Gamma| && (30)\\
&= \rho(q,p;t)|d\Gamma|\,. && (31)
\end{aligned}$$

You have the same number of neighbors as before. Thus, the probability density is constant along a trajectory.

B. Multi-particle

Now generalize to the N-particle, 1-dimensional case. Start at initial phase space coordinates $(q_1,\ldots,q_N,p_1,\ldots,p_N)$, with a small $2N$-dimensional box $d\Gamma$ of volume $|d\Gamma| = dq_1\ldots dp_N$. Under an infinitesimal step of the Hamiltonian

dynamics, the box becomes $d\Gamma'$ with volume related to $|d\Gamma|$ by the relation

$$|d\Gamma'| = \det \mathcal{J} |d\Gamma| \ , \qquad \text{where } \mathcal{J} = \mathbf{1}_{2N} + \Delta t \begin{pmatrix} \frac{\partial^2 H}{\partial p_1 \partial q_1} & \frac{\partial^2 H}{\partial p_1 q_2} & \cdots & \frac{\partial^2 H}{\partial p_1 \partial p_N} \\ \frac{\partial^2 H}{\partial p_2 q_1} & \frac{\partial^2 H}{\partial p_2 \partial q_2} & \cdots & \frac{\partial^2 H}{\partial p_2 \partial p_N} \\ & & \ddots & \\ \vdots & \vdots & -\frac{\partial^2 H}{\partial q_1 \partial p_1} & \\ & & & \ddots \\ -\frac{\partial^2 H}{\partial q_N \partial q_1} & -\frac{\partial^2 H}{\partial q_N \partial q_2} & & -\frac{\partial^2 H}{\partial q_N \partial p_N} \end{pmatrix} \equiv \mathbf{1}_{2N} + \Delta t (\delta \mathcal{J}) \ . \tag{32}$$

To calculate this, we can use the matrix identity:

$$\det \left(e^{\Delta t (\delta \mathcal{J})} \right) = e^{\mathrm{tr} \, (\Delta t (\delta \mathcal{J}))} \tag{33}$$

$$\implies \det(\mathbf{1}_{2N} + \Delta t (\delta \mathcal{J})) + \mathcal{O}(\Delta t^2) = 1 + \Delta t \ \mathrm{tr} \, (\delta \mathcal{J}) + \mathcal{O}(\Delta t^2) = 1 + \Delta t \sum_{i=1}^{N} \frac{\partial^2 H}{\partial p_i \partial q_i} - \Delta t \sum_{i=1}^{N} \frac{\partial^2 H}{\partial q_i \partial p_i} = 1 \ . \tag{34}$$

Thus the phase-space volume is conserved in the multi-particle case, as well. The generalization to higher dimensions (e.g. the $6N$-dimensional phase space of 3-dimensional, N-particle dynamics) is straightforward.

C. When it fails

Liouville's theorem isn't always true, e.g. for dissipative dynamics. For example, consider a damped harmonic oscillator with friction γ:

$$\dot{q} = p/m \tag{35}$$

$$\dot{p} = -kq - \gamma p \ . \tag{36}$$

The Jacobian for the phase space volume change is then

$$\mathcal{J} = \mathbf{1} + \Delta t \begin{pmatrix} 0 & 1/m \\ -q & -\gamma \end{pmatrix} \quad \implies \quad \det \mathcal{J} = 1 - \Delta t \gamma + \mathcal{O}(\Delta t^2) \ . \tag{37}$$

This, phase space volume shrinks at a rate of $d|d\Gamma|/dt = -\gamma |d\Gamma|$, causing the probability density to exponentially localize to $(q, p) = (0, 0)$.