

# Data science and Statistical Physics

Vipul Periwal

NIH/NIDDK/Laboratory of Biological Modeling

# Lecture 1

- 1.1 Inverse problem in statistical physics  $\geq$  data science?
- 1.2 What are the assumptions we make in statistical physics?
- 1.3 Where is information theory in all this?
- 1.4 What does large deviations mean?

## Maximum likelihood and why we need models:

If we observe frequencies  $f_i$  for specific events, then the standard way to find the probabilities of these events is to maximize  $ML = \prod p_i^{f_i}$ , subject to  $\sum_i p_i = 1$

$$\partial_{p_i} (\ln ML + \lambda(1 - \sum_i p_i)) = \frac{f_i}{p_i} - \lambda = 0$$

$$\sum_i p_i = \frac{1}{\lambda} \sum_i f_i = 1 \implies \lambda = 1 \implies f_i = p_i$$

This is a so-called trivial solution because it says that *the probability of observing any new event is just the frequency with which that event was observed already.*

Our aim is to make better predictions from observations.

- Need models for calculating probabilities  $\leftrightarrow$  frequencies
- Need way to evaluate model performance
- Need framework to compare models

Maximum likelihood and information theory

$$\ln ML = \sum_i f_i \ln p_i = \sum_i f_i \ln \frac{p_i}{f_i} + \sum_i f_i \ln f_i$$

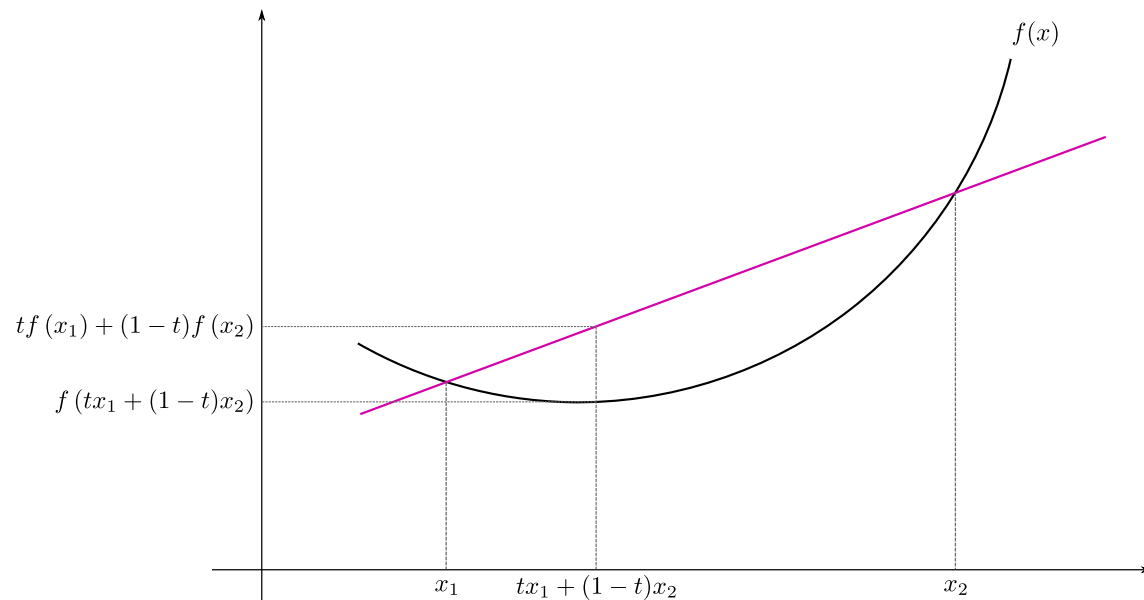
$\sum_i f_i \ln f_i$  is a constant, depending only on the data.

So we need to focus on maximizing  $\sum_i f_i \ln \frac{p_i}{f_i}$



## Convex functions:

$$f(x_1 t + x_2(1 - t)) \leq t f(x_1) + (1 - t) f(x_2) \quad \forall t \in [0, 1]$$



$$f\left(\sum_i x_i t_i\right) \leq \sum_i t_i f(x_i) \text{ if } \sum_i t_i = 1, \quad \forall t_i \geq 0$$

The general form we will use sometimes is called Jensen's inequality:

$$f(E[X]) \leq E[f(X)]$$

for any convex function  $f$  and any random variable  $X$ .

Proof:  $f(\sum_i x_i t_i) \leq \sum_i t_i f(x_i)$  if  $\sum_i t_i = 1, \forall t_i \geq 0$  by induction and limits.

Some notation: Expectation value of  $\mathcal{O} \equiv E(\mathcal{O}) \equiv \sum_A t_A \mathcal{O}(A)$  for  $t_A \geq 0, \sum_A t_A = 1$

$\sum_i f_i \ln \frac{p_i}{f_i}$  looks very similar to what we just saw for convex functions because  $\sum_i f_i = 1$  and  $f_i \geq 0$  and  $-\ln(x)$  is convex!

So maximizing  $\sum_i f_i \ln \frac{p_i}{f_i}$  is the same as minimizing

$$D_{KL}(f||p) \equiv \sum_i f_i \ln \frac{f_i}{p_i}$$

This is the Kullback-Leibler divergence, a measure of how similar two probability distributions are but it is NOT symmetric in the two arguments!!

$$\text{Convexity} \implies D_{KL}(f||p) \geq -\ln \sum_i f_i \frac{p_i}{f_i} = -\ln \sum_{i:f_i>0} p_i \geq 0$$

$$f\left(\sum_i x_i t_i\right) \leq \sum_i t_i f(x_i) \text{ if } \sum_i t_i = 1 \ \forall t_i \geq 0$$

To recap: We have found that we want to minimize  $D_{KL}(f||p) \equiv \sum_i f_i \ln \frac{f_i}{p_i}$

and we need some sort of model for the probabilities.

Statistical physics is all about probabilities following Boltzmann. A physical system with fixed energy will occupy all accessible states with equal probability and this is encoded in maximizing

$$S \equiv - \sum_i p_i \ln p_i$$

If we do a Legendre transform to a more convenient intensive variable, we want to maximize

$$S \equiv - \sum_i p_i \ln p_i - \beta \sum_i p_i E_i + \lambda (\sum_i p_i - 1)$$

$$S \equiv - \sum_i p_i \ln p_i - \beta (\sum_i p_i E_i - U) + \lambda (\sum_i p_i - 1)$$

$$\Rightarrow \frac{\partial S}{\partial p_i} = \lambda - \ln p_i - 1 - \beta E_i = 0$$

$$\frac{\partial S}{\partial \lambda} = \sum_i p_i - 1 = 0$$

$$\frac{\partial S}{\partial \beta} = \sum_i p_i E_i - U = 0$$

**Note:**  $S = \beta(U - F)$

$$\Rightarrow p_i = \frac{\exp(-\beta E_i)}{Z}$$

$$\frac{\partial(\beta F)}{\partial \beta} = U(\beta)$$

$$Z \equiv \exp(-\beta F) \equiv \sum_i \exp(-\beta E_i)$$

So now we start to see what form of probabilities we could use in our inference. We expect in this analogy that **higher energy states will have lower probability** and that is why we observe them with lower frequency.

How do we know that the appropriate probability distribution we are trying to infer is of this exponential form??

We could always write  $p_i = \exp(\ln p_i)$

BUT we will see that there are many calculations that we want to do that do not work unless  $\ln p_i = \theta \cdot \vec{O}_i + f(\theta)$

More formally, we will be able to investigate models where the probability distribution that we are trying to infer belongs to the **exponential family**.

Exponential family distributions:

- normal
- beta
- Poisson
- exponential
- Dirichlet
- gamma
- Bernoulli
- Wishart
- geometric
- multinomial (with fixed # of trials)


Must take the form

$$p(x|\theta) = h(x)g(\theta) \exp(\theta \cdot f(x))$$


observation  
variables



parameters we  
want to determine  
from data



set of functions of  
the observation  
variables



Some examples:

$$\sigma \in \{-1, 1\}$$

$$p(\sigma|\theta) = \frac{\exp(\sigma\theta)}{2 \cosh \theta}$$

$$\sigma_i \in \{0, 1\}, i = 0, 1$$

$$p(\sigma_0, \sigma_1 | \vec{\theta} = (\theta_0, \theta_1, \theta_{01})) = \frac{e^{\vec{\theta} \cdot (\sigma_0, \sigma_1, \sigma_0 \sigma_1)}}{1 + e^{\theta_0} + e^{\theta_1} + e^{\theta_0 + \theta_1 + \theta_{01}}}$$

$$r \geq 0, r \in \mathbf{Z}$$

$$p(r|\lambda) = \frac{\lambda^r}{r!} e^{-\lambda} = \frac{e^{-\lambda}}{r!} e^{r \ln \lambda}$$

`Natural parameter':  $\ln \lambda$



We will always work with natural model parameters that we want to infer because of **convexity**.

Generating function of correlations:  $Z \equiv \sum e^{\theta \cdot \mathcal{O}}$

$$\frac{\partial \ln Z}{\partial \theta_i} = \sum \mathcal{O}_i \frac{e^{\theta \cdot \mathcal{O}}}{Z} = E(\mathcal{O}_i)$$

$$\frac{\partial^2 \ln Z}{\partial \theta_j \partial \theta_i} = E(\mathcal{O}_i \mathcal{O}_j) - E(\mathcal{O}_i)E(\mathcal{O}_j) \equiv E_c(\mathcal{O}_i \mathcal{O}_j) \equiv C_{ij}$$

$$\sum_{i,j} v_i v_j C_{ij} = E((v \cdot \mathcal{O} - E(v \cdot \mathcal{O}))(v \cdot \mathcal{O} - E(v \cdot \mathcal{O}))) \geq 0$$

So the **log of Z is convex** as a function of natural parameters!

Correlation functions  $\Leftrightarrow$  moments in statistics

Connected correlation functions  $\Leftrightarrow$  cumulants in statistics

So the generating function of moments is  $Z(\theta)$

and the generating function of cumulants is  $\ln Z(\theta) \equiv W(\theta)$

Remember that  $\theta$  denotes the vector of natural parameters!

$$E_c(\mathcal{O}_i \mathcal{O}_j) = E((\mathcal{O}_i - E(\mathcal{O}_i))(\mathcal{O}_j - E(\mathcal{O}_j))) \Leftarrow \text{Why??}$$

Let's see how this generating function is going to help us find parameters from data. Going back to the Kullback-Leibler divergence, we want to make the model probability distribution as close to the observed frequencies as possible.

$$D_{KL}(f||p) = \sum_A f_A (\ln f_A - \ln p_A) = \sum_A f_A (\ln f_A - \theta \cdot \mathcal{O}(A) + \ln Z)$$

Sum over  
observations

$$\frac{\partial D_{KL}(f||p)}{\partial \theta_i} = E(\mathcal{O}_i) - \sum_A f_A \mathcal{O}_i(A)$$

$$\frac{\partial D_{KL}(f||p)}{\partial \theta_i} = E(\mathcal{O}_i) - E_f(\mathcal{O}_i)$$

Expectation in data

Expectation in model,  
difficult to compute  
because it requires a sum  
over all configurations!

Use this for gradient  
descent to find  
parameters.

## Statistical physics from the perspective of probability theory

1. Macroscopic description  $\Leftrightarrow$  Law of large numbers
2. Behavior of fluctuations  $\Leftrightarrow$  Central limit theorem
3. Probability of rare events  $\Leftrightarrow$  Large deviation principles
4. Probability of empirical pdf  $\Leftrightarrow$  Level 2 LDP

## Law of large numbers

$$\frac{1}{N} \sum_{i=0}^{N-1} x_i \asymp \mu \quad (N \rightarrow \infty)$$

Chebyshev's theorem:

For a random variable with a finite expected value and a finite non-zero variance,

$$Pr(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Proof:

$$\begin{aligned} \sigma^2 &= E(|X - \mu|^2) \\ &= E(|X - \mu|^2 | |X - \mu| \leq k\sigma) Pr(|X - \mu| \leq k\sigma) + E(|X - \mu|^2 | |X - \mu| > k\sigma) Pr(|X - \mu| > k\sigma) \\ &\geq 0 \times Pr(|X - \mu| \leq k\sigma) + (k\sigma)^2 Pr(|X - \mu| > k\sigma) \end{aligned}$$

## Central limit theorem CLT

Under quite general circumstances, the *appropriately normalized sum of independent* random variables (RVs), the distribution of the sum will tend (in the limit) to the *normal distribution*. The reason this distribution is called the *Gaussian* is because (as with so much else) Gauss figured this out first.

$$N(x|\mu, \sigma) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

The CLT lies at the heart of many proofs in statistics for the asymptotic behavior of inference algorithms, but keep in mind that it is a *theorem* and it comes with *hypotheses* so you can't blindly assume that it will always hold for real data!

## Statement of the CLT

$$\lim_{n \uparrow \infty} Y_n \equiv \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \sim N(\cdot | 0, \sigma) \quad \text{if } E(X_i) = 0 \text{ and } Var(X_i) = \sigma^2 \quad \forall i, X_i \text{ i.i.d.}$$

This is *NOT* uniform convergence but only convergence in distribution. Why is that important? The tails of the distribution converge more slowly than the center. So the CLT is really telling you about *moderate* deviations from the mean.

What's behind the CLT?

$$X_1 \sim N(\cdot|\mu_1, \sigma_1) \text{ and } X_2 \sim N(\cdot|\mu_2, \sigma_2) \implies$$

$$X_1 + X_2 \sim N(\cdot|\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$$

$$X \sim N(\cdot|\mu, \sigma) \implies cX \sim N(\cdot|c\mu, c\sigma)$$

This second property is essentially a defining property of the normal distribution. From

$$Y_n \equiv \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i$$

we see that the limiting distribution will be normal, so really the question is entirely one of *convergence*.



## Moment generating function

$$M_X(t) \equiv E(\exp(tX))$$

Expand in formal powers of  $t$  and you get coefficients that are the expectations of powers of  $X$ , by definition, the moments.

## Useful identities

$$M_{X+Y}(t) = M_X(t)M_Y(t) \text{ if } X, Y \text{ independent.}$$

$$M_{cX}(t) \equiv E(\exp(ctX)) = M_X(ct)$$

## Cumulant generating function

$$K_X(t) \equiv \log M_X(t)$$

$$K_{X+Y}(t) = K_X(t) + K_Y(t), \quad K_{cX}(t) = K_X(ct)$$

$$K_X(t=0) = 0$$

$$K'_X(t=0) = E(X)$$

$$K''_X(t=0) = E(X^2) - E(X)^2 \equiv \text{Var}(X)$$

$$K_X^{(m)}(t=0) = m^{\text{th}} \text{ cumulant}$$

## Cumulant scaling

If  $K_X^{(m)} < C < \infty \forall m$  then  $K^{(m)}(Y_n) \leq \frac{nC}{n^{m/2}} \rightarrow 0$  for  $m > 2$

This is really the proof. It assumes that the RV has been shifted so that the expectation value of each  $X$  is 0. In other words, the cumulants of the sums tend to the cumulants of a normal distribution! We left out the Levy continuity theorem which is needed to show that the implicit assumption that the moment generating function converges doesn't matter.

## A couple of examples ...

### Bernoulli RV $\rightarrow$ Binomial RV

$$K_{\text{Ber}(p)}(t) = \ln(M_{\text{Ber}(p)}(t)) = \ln((1-p) + p \exp(t)) = tp + \frac{t^2}{2}p(1-p) + \dots$$

$$K_{\text{Bin}(n,p)}(t) = \ln(M_{\text{Bin}(n,p)}(t)) = n \ln M_{\text{Ber}(p)}(t) \xrightarrow{n \uparrow \infty} (np)(\exp t - 1)$$

### Poisson

$$\text{Poisson}(k|\lambda) \equiv \exp(-\lambda) \frac{\lambda^k}{k!}$$

$$K_{\text{Poisson}}(t) = \ln(\exp(-\lambda) \sum_{k=0}^{\infty} \frac{\lambda^k \exp(tk)}{k!}) = -\lambda + \lambda \exp(t)$$

$$\lambda \leftrightarrow np$$


## Moving on to more control of rare events ...

Suppose we have random bits taking values 0 or 1. We want to calculate the probability that the mean value of the bits observed is a certain value.

$$R_n \equiv \frac{1}{n} \sum_{i=0}^{n-1} b_i \quad \text{We want to calculate } P(R_n = r).$$

Every possible vector of  $n$  bits is equally likely so we want to sum over all possible  $n$  bit vectors with the correct sum specified by  $r$ .

$$\sum_{b: R_n(b)=r} 2^{-n} = 2^{-n} \frac{n!}{(rn)!((1-r)n)!}$$

Stirling's approximation

$$n! \approx n^n \exp(-n)$$

$$(rn)! \approx \exp(rn \ln(rn)) \exp(-rn)$$

So we find our first Large Deviations result

$$P(R_n = r) \approx \exp(-nI(r))$$

$$I(r) = \ln 2 + r \ln r + (1 - r) \ln(1 - r)$$

Notice that  $I(r)$  is convex and has a unique zero at  $r = 1/2$ .

 **Rate function** (binary cross-entropy form)

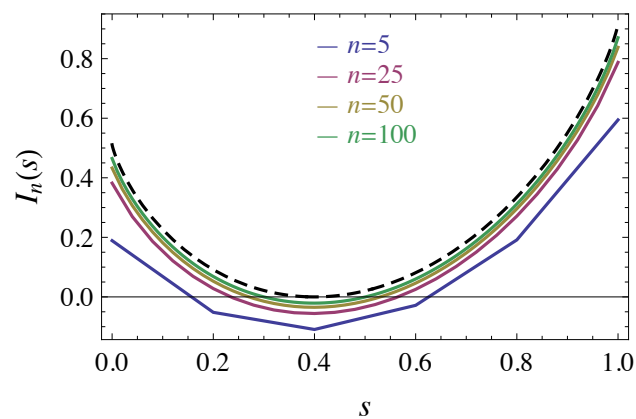
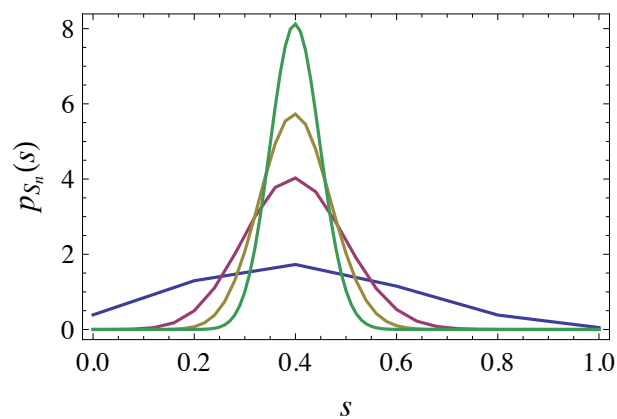
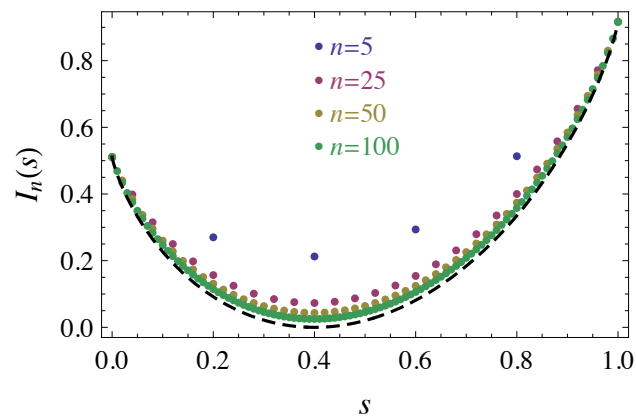
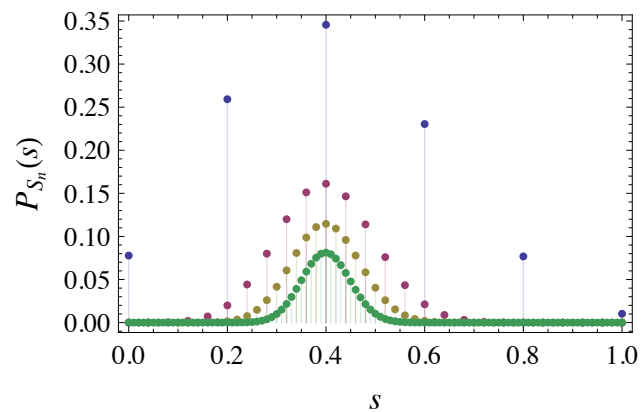
$\{0, 1\}$

$$P(X_i = 0) = 1 - \alpha \text{ and } P(X_i = 1) = \alpha$$

$$I(s) = s \ln \frac{s}{\alpha} + (1 - s) \ln \frac{1 - s}{1 - \alpha},$$

$$P_{S_n}(s) \approx e^{-nI(s)}$$

$$s \in [0, 1]$$



From H. Touchette, The large deviation approach to statistical physics

Another example: i.i.d. normal variables

$$P(Y_n = s) = \int_{\vec{x} \in \mathbf{R}^n} \delta(Y_n(\vec{x}) - s) \prod_i dx_i N(x_i | \mu, \sigma)$$

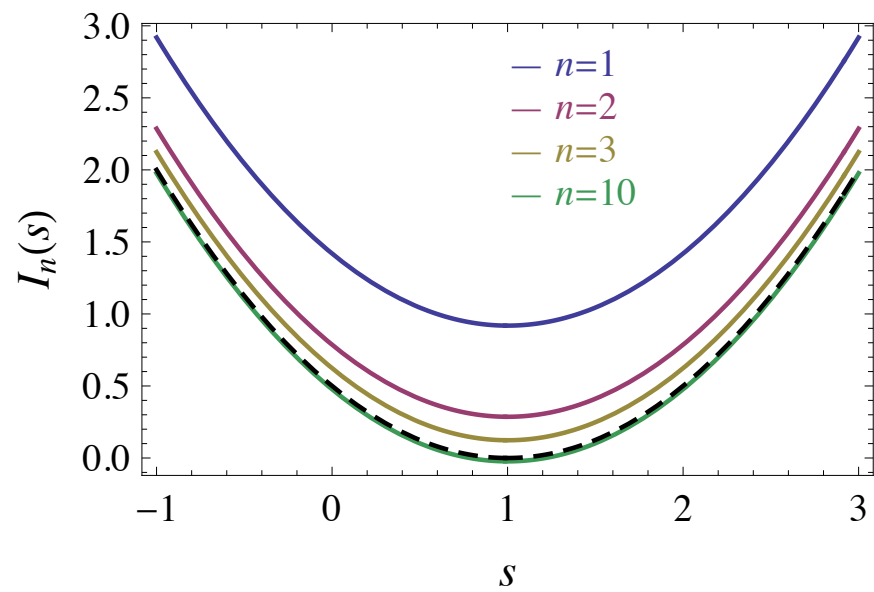
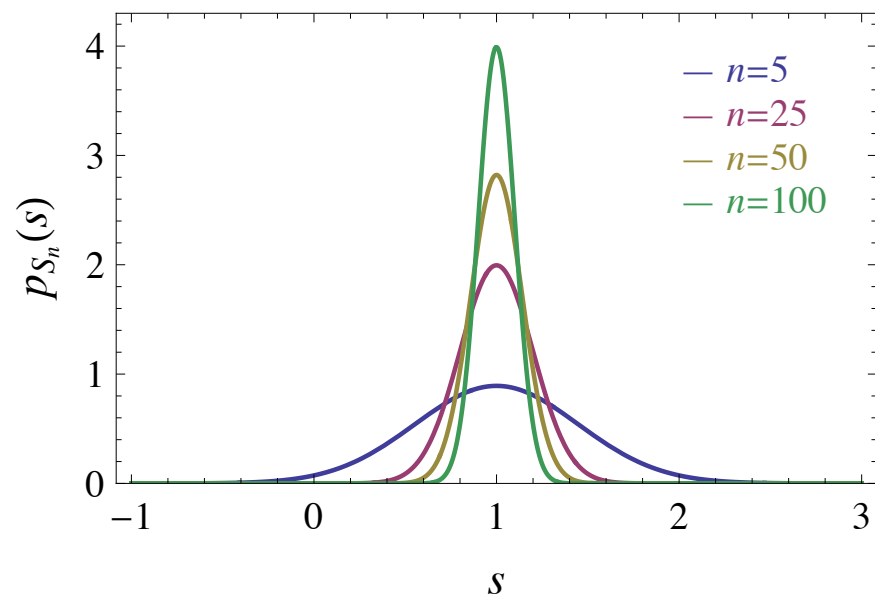
Subleading term  $\Rightarrow \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n(s - \mu)^2}{2\sigma^2}\right)$

So we find our second Large Deviations result

$$I(s) = \frac{(s - \mu)^2}{2\sigma^2}$$

Again convex and with a unique zero.





From H. Touchette, The large deviation approach to statistical physics

Yet another example: i.i.d. exponential variables

$$X \sim \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \text{ for } \mu > 0$$

$$P(Y_n = s) = \left(\frac{1}{\mu} \exp\left(-\frac{s}{\mu}\right)\right)^n \int_0^\infty \prod dx_i \delta\left(s - \frac{1}{n} \sum x_i\right)$$

So, after doing the integral and using Stirling again, we get

$$I(s) \approx \frac{s}{\mu} - 1 - \ln \frac{s}{\mu}, \quad s > 0$$

Again convex and with a unique zero at  $s = \mu$

Now some theory ...

A large deviation principle amounts to

Some probability  $P_n \approx \exp(-nI)$  as  $n \uparrow \infty$ , with  $I \geq 0$ .

More formally:

With  $A_n$  a family of RVs and  $B$  a set,

$$\lim_{n \uparrow \infty} -\frac{1}{n} \ln P(A_n \in B) = I_B$$

$$I_B \equiv \text{rate.}$$

To connect to the rate functions we have calculated,

$$P(Y_n \in [s, s + ds]) \approx \exp(-nI(s))ds$$

Now we turn to more direct ways to calculate rate functions.

## Gärtner-Ellis Theorem

Suppose we have found a rate function,  $I(s)$ . Then we have

$$P(Y_n \in [a, a + da]) \approx \exp(-nI(a))da.$$

So

$$\langle \exp(tY_n) \rangle \approx \int \exp(ta - nI(a))da.$$

Setting  $t \equiv kn$  for some real number  $k$ , we have

$$\langle \exp(tY_n) \rangle \approx \int \exp n(ka - I(a))da.$$

Laplace's approximation (or saddle-point) then says that

$$\left\langle \exp(tY_n) \right\rangle \approx \exp n \sup_a (ka - I(a)).$$

In other words,

$$\lambda(k) \equiv \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left\langle \exp(knY_n) \right\rangle \approx \sup_a (ka - I(a)).$$

What is this telling us??

The cumulant (after scaling!) is the Legendre-Fenchel transform of the rate function!

# The Legendre-Fenchel transform

For a convex function  $f(t)$ , define a new function by

$$g(s) \equiv \sup_t (ts - f(t)).$$

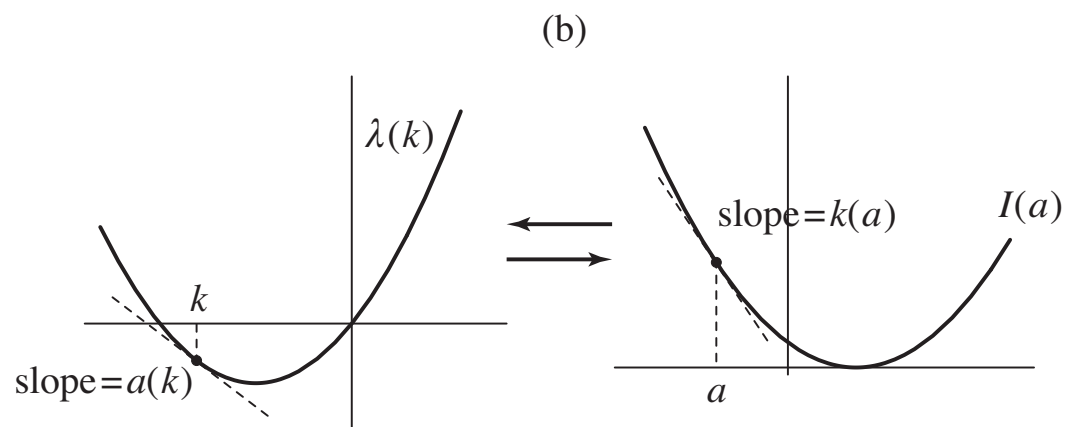
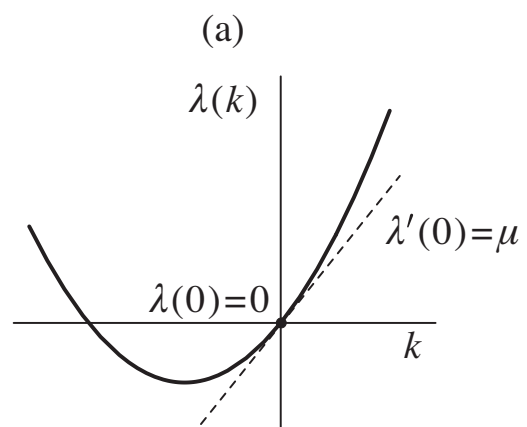
Then  $g(s)$  is also convex and we have a duality:

$$f(t) \equiv \sup_s (ts - g(s)).$$

This is the more general form of the usual physics Legendre transform:

$$f(t) + g(s) = ts,$$

which is only valid when  $f, g$  are differentiable.



From H. Touchette, The large deviation approach to statistical physics



## Reminder of how Legendre transforms work

Pick a variable,  $t$  or  $s$ . Then we regard all appearances of the other variable as an implicit function of the chosen variable. Treating  $s$  as a fixed parameter, we get

$$\partial_t f = s.$$

This defines

$$s(t) = \partial_t f, \quad \text{and } \textit{vice versa} \quad t(s) = \partial_s g(s).$$

Now we have the very important relation:

$$\partial_t s(t) = \partial_t \partial_t f, \quad \text{and} \quad \partial_s t(s) = \partial_s \partial_s g.$$

But this gives us

$$\partial_t \partial_t f = (\partial_s \partial_s g)^{-1}.$$

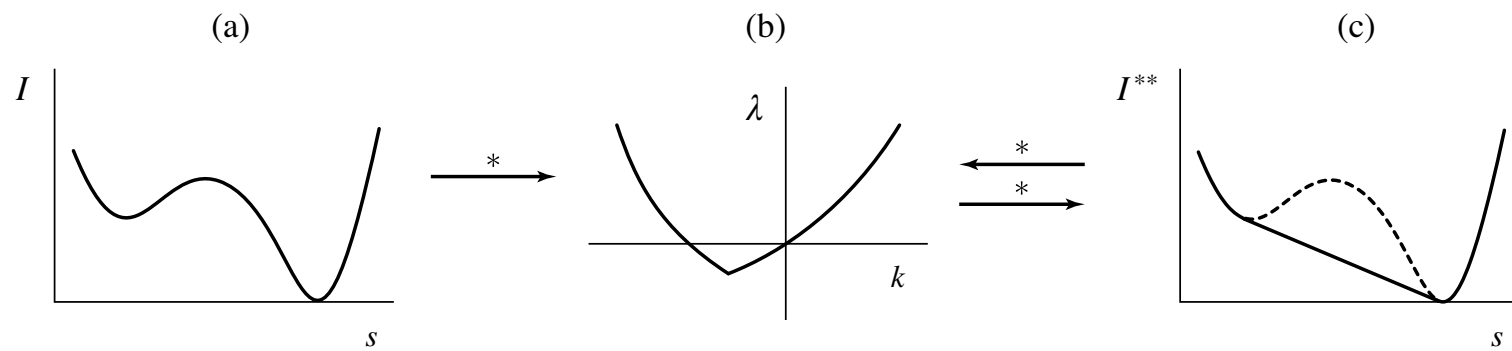
## How do we use this abstract nonsense?

Suppose we have a cumulant generating function for an RV  $X$  :  $K_X(t) = \ln M_X(t)$ . Then

$$E(X)(t) = \partial_t K_X(t) = \frac{E(X \exp(Xt))}{E(\exp(Xt))}.$$

In the language of Legendre transforms,  $E(X)(t) \equiv X_c(t)$  is the analog of  $s(t)$ . So by thinking in terms of the Legendre transform, we see that the large deviation rate function, which is a function of the expectation value, is naturally the Legendre transform of the cumulant generating function.

**WARNING:** What if the rate function is not convex? The cumulant generating function is always convex (Why?). The Legendre transform of a non-convex function is convex, but the double Legendre transform does not give back the original non-convex function!!



From H. Touchette, The large deviation approach to statistical physics

## Cramér's Theorem

This is where the theory of large deviations actually started. Here, we approach it as an application of the Gärtner-Ellis Theorem. Suppose  $X_i$  are i.i.d. RVs. Then the scaled cumulant generating function is

$$\lambda(k) = \lim_{n \uparrow \infty} \frac{1}{n} \ln \left\langle \exp\left(k \sum_i X_i\right) \right\rangle = \lim_{n \uparrow \infty} \frac{1}{n} \ln \prod_i \left\langle \exp(k X_i) \right\rangle = \ln \left\langle \exp(k X) \right\rangle.$$

In other words, the cumulant generating function of any of the variables *is* the scaled cumulant function. So we can get the rate function directly from this cumulant function.

We will use this result over and over.

## Example again: i.i.d. normal variables

$$\lambda(k) \equiv \ln \langle \exp(kX) \rangle = \ln \left( \frac{\int \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \exp(kx)}{\int \exp(-\frac{(x-\mu)^2}{2\sigma^2})} \right) = \mu k + \frac{\sigma^2 k^2}{2},$$

so we need the Legendre transform of  $\mu k + \frac{\sigma^2 k^2}{2}$ . As  $\lambda(k)$  is differentiable,

$$\partial_k \lambda(k) = s(k) = \mu + \sigma^2 k$$

so

$$k(s) = \frac{s - \mu}{\sigma^2}.$$

Therefore

$$I(s) = k(s)s - \lambda(k(s)) = \frac{s - \mu}{\sigma^2} s - \mu \frac{s - \mu}{\sigma^2} - \frac{1}{2\sigma^2} (s - \mu)^2 = \frac{1}{2\sigma^2} (s - \mu)^2$$

as we found before.

## Again with Cramer: i.i.d. exponential variables

$$\lambda(k) \equiv \ln \langle \exp(kX) \rangle = -\ln(1 - k\mu),$$

so again  $\lambda(k)$  is differentiable, and we have

$$\partial_k \lambda(k) = s(k) = \frac{\mu}{1 - k\mu}.$$

Now

$$k(s) = \frac{s - \mu}{s\mu}.$$

Therefore

$$I(s) = k(s)s - \lambda(k(s)) = \frac{s}{\mu} - 1 + \ln \left( 1 - \frac{s - \mu}{s} \right) = \frac{s}{\mu} - 1 + \ln \left( \frac{\mu}{s} \right)$$

as we found before.

# Sanov's Theorem

All our results work perfectly fine for vector RVs. Sanov's theorem says that they also work for functions that are RVs. One case in particular is exactly what data science is all about: Determining the probability density underlying a set of empirical observations. Define an empirical probability density  $L_n$  :

$$L_n(x) \equiv \frac{1}{n} \sum_i \delta(\sigma_i - x), \text{ where } \{\sigma_i\} = \text{observations.}$$

$L_n(x)$  is a function that is an RV. Now

$$\lambda(k) = \ln \int dx \, \rho(x) \exp(k(x)) \equiv \ln \langle \exp k(X) \rangle.$$

Notice that  $k$  is now a function! With the same arguments as before, we find a rate function which measures the probability of any given probability density  $\mu$  :

$$I(\mu|\rho) = \int dx \, \mu(x) \ln \frac{\mu(x)}{\rho(x)}.$$

You should recognize our old friend, the Kullback-Leibler divergence, here.

## Inverse Sanov's Theorem

To be honest, Sanov's theorem doesn't help us directly because we don't know  $\rho$ , the true probability density. What we want is a data-driven version that tells us about convergence to the true density but based only on the observed data. The approach usually to prove an Inverse Sanov Theorem is Bayesian:

$$P(\rho|L_n) = \frac{P(L_n|\rho)P(\rho)}{P(L_n)},$$

where  $\rho$  is a possible model for the true distribution, and Sanov's theorem is used for calculating  $P(L_n|\rho)$ . This requires a little more mathematical effort. We may return to it later.



# Varadhan's Theorem

$$\lambda(f) = \lim_{n \uparrow \infty} \frac{1}{n} \langle \exp nk f(A_n) \rangle = \sup_a (f(a) - I(a))$$

for any continuous bounded function  $f$ . This is not just a trivial generalization. It holds in much more general circumstances than the Laplace saddle-point argument that we gave to justify when  $f(a) = ka$ .

# Contraction principle

Suppose  $B_n$  is some other RV, and  $B_n = h(A_n)$ , where  $h$  may be many-to-one. Then a rate function for  $A_n$  implies a rate function for  $B_n$  because

$$I_B(b) = \inf_{a:h(a)=b} I_A(a).$$

Why is this true?

The LEAST improbable value of  $a$  for which  $h(a) = b$ !

What happens if there is no such value of  $a$ ? Then  $I_B(b) = \infty$ , because such a value of  $b$  is never observed.

# Warning example!

Suppose

$$p(x) = \frac{C}{(|x| + a)^\beta}, \quad \beta > 3$$

for  $x$  real. Now  $Var(X)$  is finite, so the CLT holds. But what is  $I(s)$ ?

$$\lambda(k) = \infty \quad \forall k \neq 0.$$

Therefore  $I(s) = 0$ .

In general, we can use Cramér's theorem *locally* at values where  $\lambda(k)$  is differentiable.

Suppose the RV is a ‘stuck’ coin, so it’s either  $1, 1, 1, \dots$  or  $-1, -1, -1, \dots$ . Then

$$p(Y_n = y) = \frac{1}{2} (\delta(y + 1) + \delta(y - 1)).$$

What is  $I(s)$ ? First, we calculate

$$\lambda(k) = \lim_{n \uparrow \infty} \frac{1}{n} \ln \cosh(nk) = |k|.$$

On the other hand, we can directly see that

$$I(s) = 0 \text{ for } s = \pm 1, \quad \text{and } I(s) = \infty \text{ everywhere else.}$$

This is non-convex! The Legendre-Fenchel transform is

$$I_{\text{LF}}(s) = \sup_k (ks - |k|).$$

For  $s < -1$ , or  $s > 1$ , the supremum gives  $I_{\text{LF}}(s) = \infty$ , but for  $s \in [-1, 1]$ , we have  $I_{\text{LF}}(s) = 0$ !

Suppose  $Y_n = Z + \frac{1}{n} \sum_i X_i$ , where  $X_i$  are i.i.d. normal RVs and  $Z = \pm 1$  with probability  $\frac{1}{2}$ . We already know

$$P(Y_n = s | Z = \pm 1) \approx \exp(-nI_{\pm}(s)), \quad \text{with } I_{\pm}(s) = \frac{(s \mp 1)^2}{2\sigma^2}.$$

so we get (summing over the probabilities of the  $Z$  values)

$$P(Y_n = s) \approx \exp(-nI(s)), \quad I(s) = \min(I_+(s), I_-(s)).$$

Now we have

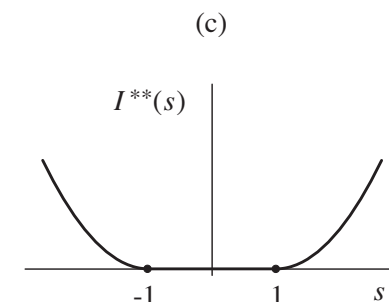
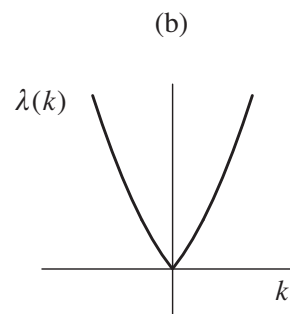
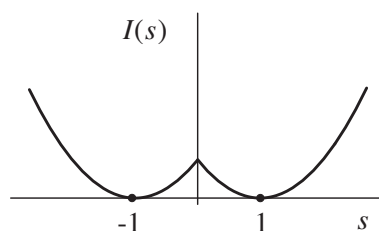
$$\lambda(k) = |k| + \frac{k^2}{2},$$

and like the previous example

$$I_{\text{LF}}(s) = 0 \quad \text{for } s \in [-1, 1],$$

(a)

with  $I_{\text{LF}}(s) = I_{\pm}(s)$  for  $\pm s > 1$ .



**From H. Touchette, The large deviation approach to statistical physics**

# Another type of random variable

Consider sequences  $\omega \equiv (\omega_1, \omega_2, \dots, \omega_n)$ .  $\omega_i$  could be random i.i.d. or a Markov chain. Then the probability of a sequence  $P_n(\omega)$  is an RV, and we can calculate the rate of

$$A_n(\omega) \equiv -\frac{1}{n} \ln P_n(\omega).$$

If random i.i.d.,

$$\lambda(k) = \lim_{n \uparrow \infty} \frac{1}{n} \ln \langle P_n^{-k} \rangle = \lim_{n \uparrow \infty} \frac{1}{n} \ln \sum_{\omega} P_n(\omega)^{1-k} = \sum_j P(\omega_i = j)^{1-k}.$$

Suppose  $I(a)$  is convex and has a unique minimum,  $a^*$ . Then

$$\lim_{n \uparrow \infty} \langle A_n \rangle = a^*,$$

but also

$$\lim_{n \uparrow \infty} A_n = a^* \text{ almost surely.}$$

The first implies that the mean Boltzmann-Gibbs-Shannon entropy

$$H_n \equiv - \sum_{\omega} P_n(\omega) \ln P_n(\omega) \rightarrow a^*,$$

which is the Kolmogorov-Sinai entropy or the entropy rate. The second is the Asymptotic Equipartition Theorem or the Shannon-McMillan-Breiman theorem, and it says that most of the probability is in sequences with  $P_n(\omega) \approx \exp(-na^*)$ . So how many such *typical* sequences are there?  $\exp(na^*)$ .

# Large Deviations and Equilibrium Statistical Physics

Microstates  $\leftrightarrow \omega = (\omega_1, \omega_2, \dots, \omega_n)$

Macrostate  $\leftrightarrow f_n(\omega)$

*A priori* distribution on the space of microstates  $P(d\omega)$

Mean energy per particle  $h_n(\omega) \equiv H_n(\omega)/n$ .

Thermodynamic limit  $n \rightarrow \infty$

Basic idea:

Observed macroscopic variables concentrate at minima of their rate functions



# Entropy

$$\Omega(h_n \in du) = \int_{\omega: h_n(\omega) \in du} d\omega$$

assuming a uniform measure on the space of microstates.

Define a rate function

$$I(u) = \lim_{n \uparrow \infty} -\frac{1}{n} \ln P(h_n \in du)$$

assuming a uniform it a priori measure  $d\omega/|\Lambda|^n$ .

Then

$$I(u) = \ln |\Lambda| - \lim_{n \uparrow \infty} \frac{1}{n} \ln \Omega(h_n \in du).$$

Rate function = – Entropy up to log of ‘volume’

$$P(h_n \in \mathrm{d}u) \propto \Omega(h_n \in \mathrm{d}u)$$

so with

$$Z_n(\beta) = \int_{\Lambda^n} \mathrm{d}\omega \exp(-\beta H_n(\omega))$$

we have

$$\lambda(k) \equiv \lim_{n \uparrow \infty} \frac{1}{n} \ln \langle \exp(k H_n) \rangle = \lim_{n \uparrow \infty} \frac{1}{n} \ln Z_n(\beta) \Big|_{\beta=-k} - \ln |\Lambda|.$$

It follows that the Massieu potential  $\phi(\beta) = -\lambda(k = -\beta)$  is a concave function of  $\beta$  (we normalize  $|\Lambda| = 1$ ).

Using Varadhan's theorem and the Gärtner-Ellis theorem, we see our usual dualities:

$$\phi(\beta) = \inf_u (\beta u - s(u))$$

and

$$s(u) = \inf_{\beta} (\beta u - \phi(\beta)).$$

(inf?? Why?)