Information Theory & Statistical Physics

Lecture: Prof. Dr. Johannes Berg
Exercises: Stephan Kleinbölting **Sheet 3**

Due date: ***Fri, 02.06.17 12:00*** *To be discussed on: to be announced*

Website: http://www.thp.uni-koeln.de/~skleinbo/teaching/Information2017/

If you have questions regarding the presented solutions or the accompanying Julia notebook, please do not hesitate to ask!

11 Rare events

40 pts.

Consider a system of n Ising spins $\mathbf{X} \in \{-1, +1\}^n$ that individually point up with probability f. On average the system will show a net magnetization $m := \langle M \rangle = (2f - 1)n$. We would like to explore how likely deviations from the mean are.

Specifically, consider the event

$$E = \{P \in \mathcal{P} : \sum_{x \in \pm 1} xP(x) - m > \alpha\}$$

Sanov's theorem tells us that the probability of E behaves like

$$\mathrm{P}_Q(E) \doteq \exp\left[-nD(P^*||Q)\right] \tag{1}$$

to first order in the exponent. P^* is the distribution which minimizes the Kullback-Leibler divergence to the true distribution under to the constraints set by E.

(a) *25pt* - Find P^* for an arbitrary distribution Q and show that it is given by

$$P^*(x) = Q(x) \exp(\lambda_0 + \lambda_1 x). \tag{2}$$

i.e minimize

$$D(P||Q) \quad \text{under the constraint} \quad \sum_x xP(x) \geq \alpha.$$

Determine $\lambda_{0,1}$ for $Q = (1 - f, f)$ from the constraint and normalization.

Now specialize to $f = 1/2$. You should find $\lambda_1 = \tanh^{-1}(\alpha)$.

What's the probability to observe at least 700 of 1000 spins pointing up? Could you maybe have guessed P^*?

(b) *15pt* - The *central limit theorem* ensures that the sum of iid. random variables with expectation μ and variance σ^2 approaches a normal distribution. More precisely, given iid. random variables $X_1, \dots X_n$ as stated

$$\frac{Z_n - \mu n}{\sqrt{n}} \equiv \frac{\sum_{j=1}^n X_j - \mu n}{\sqrt{n}} \to \mathcal{N}(0, \sigma^2) \text{ point-wise.}$$

In particular let $\Phi(x, \sigma^2)$ denote the CDF of a centered normal distribution with variance σ^2, then for large enough n

$$\mathrm{P}(Z_n \geq \sqrt{n}\alpha) \approx 1 - \Phi(\alpha, \sigma^2)$$

Approximate the CDF for large values of n and compare to the result from a).

Solution

(a) Let's consider the more general problem of minimizing the mutual entropy $D(P||Q)$ with respect to P under a set of m linear constraints

$$k = 1, \ldots, m: \quad \sum_x p(x) f_k(x) \geq d_k \quad \text{and} \quad \sum_x p(x) = 1.$$

First one may transform the inequality constraints to a more canonical form by setting $\tilde{f}_k = -(f_k - d_k)$. The optimization problem then reads

$$\begin{aligned} \text{minimize} \quad & \sum_x p(x) \ln \frac{p(x)}{Q(x)} \\ \text{subject to} \quad & \sum_x p(x) \tilde{f}_k(x) \leq 0, \; \sum p(x) = 1 \end{aligned}$$

If the constraints were equalities one would proceed by introducing Lagrange multipliers and look for minima of the corresponding Lagrangian. This is still a valid approach, but one has to appreciate first that the minimum is attained when the constraints are tight.

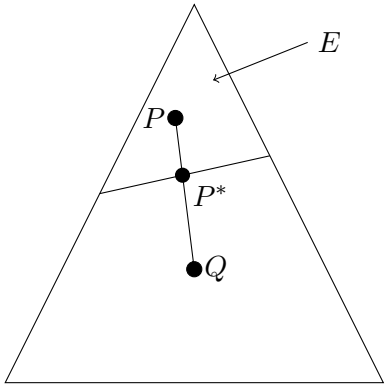

Figure 1: The space of all probability distributions $\mathcal{P}$ is a simplex. E is the event under consideration; by definition a convex subset of $\mathcal{P}$. Crucially, Q is not an element of E.

Assume therefor we had found a minimum of D in the interior of E, $P \in E^0$. Note that due to the linearity of the constraints, E is a convex set. Then consider the line segment connecting P and Q (see the figure): $(1-t)Q + tP, t \in [0,1]$. It necessarily intersects the boundary of E in a point $P^* \neq P$. Let the corresponding line parameter be $0 < t^* < 1$. The KL-divergence is a convex function (of both its arguments)

$$\begin{aligned} D(P^*||Q) &= D((1-t^*)Q + t^*P||Q) \\ &\leq (1-t^*)D(Q||Q) + t^*D(P||Q) \\ &= t^*D(P||Q) < D(P||Q) \end{aligned}$$

Therefor P is not the minimum, contradicting the assumption. Hence the minimum must be attained at the boundary, i.e when the constraints are tight.

We may thus proceed in the usual fashion and look for extrema of the Lagrangian

$$L(\{p(x)\}; \lambda_0, \cdots, \lambda_m) = \sum_x p(x) \left(\ln \frac{p(x)}{Q(x)} - \lambda_0 - \sum_{k=1}^{m} \lambda_k \tilde{f}_k(x) \right)$$

yielding

$$\forall x: \quad 0 = \ln \frac{p(x)}{Q(x)} - \lambda_0 - \sum_{k=1}^{m} \lambda_k \tilde{f}_k(x)$$

and consequently

$$p(x) = Q(x) \exp\left(\lambda_0 + \sum_{k=1}^{m} \lambda_k f_k(x)\right)$$

Again e^{λ_0} plays the role of the partition function

$$p(x) = \frac{Q(x) \exp\left(\sum_{k=1}^{m} \lambda_k f_k(x)\right)}{\sum_a Q(a) \exp\left(\sum_{k=1}^{m} \lambda_k f_k(a)\right)} = \frac{Q(x) \exp\left(\sum_{k=1}^{m} \lambda_k f_k(x)\right)}{Z}$$

In the problem at hand the only constraint is $f_1(x) = x$, which leads to the desired expression.

Now we turn to the two-state system, thus $x = \pm 1$ and $Q = (f, 1-f)$. The Lagrange multiplier λ_1 is determined by the constraint (set $y \equiv e^{\lambda_1}$)

$$\begin{aligned}\alpha = \sum_x x p(x) &= \frac{\sum_x x Q(x) \exp(\lambda_1 x)}{\sum_a Q(a) \exp(\lambda_1 a)} \\ &= \frac{fy - (1-f)/y}{fy + (1-f)/y}\end{aligned}$$

Solving for y gives

$$\lambda_1 = \ln\left(\sqrt{\frac{(1-f)(1+\alpha)}{f(1-\alpha)}}\right) = \frac{1}{2}\ln\left(\frac{(1-f)(1+\alpha)}{f(1-\alpha)}\right).$$

By virtue of the identity $\tanh^{-1}(x) = 1/2 \ln(\frac{1+x}{1-x})$ this expression reduces for $f = 1/2$ to

$$\lambda_1 = \tanh^{-1}(\alpha)$$

and

$$Z = \cosh(\lambda_1).$$

Hence the minimal mutual entropy under the constraints is given by

$$\begin{aligned}D(P^*||Q) &= \sum_x p^*(x) (\lambda_1 x - \ln Z) = \tanh^{-1}(\alpha)\langle X \rangle - \ln\cosh(\tanh^{-1}(\alpha)) \\ &= \tanh^{-1}(\alpha)\alpha - \ln\cosh(\tanh^{-1}(\alpha)) = \frac{\alpha^2}{2} + \mathcal{O}(\alpha^4)\end{aligned}$$

and by Sanov's theorem we find that

$$P(E) \doteq \exp(-n\alpha^2/2)$$

for small α.

Having 700 of 1000 spin up corresponds to $\alpha = \frac{700-300}{n} = 0.4$. The probability of observing such an event is of the order e^{-800}. One might have guessed this result by noting that the only binary distribution that has expected value of 0.7 is $P^* = (0.7, 0.3)$ yielding $D((0.7, 0.3)||(0.5, 0.5)) \approx 0.082$ in very good agreement with the previous calculation.

(b) *Note:* The term "point-wise" in the statement of the CLT is unfortunate. More precisely it must read "[...] converges in distribution", i.e the cumulative distribution functions converges point-wise

$$F_{X_n}(x) \to F_X(x) \quad \text{wherever } F_X \text{ is continuous.}$$

In any case, we may estimate

$$P(Z_n/n \geq \alpha) = P(Z_n/\sqrt{n} \geq \sqrt{n}\alpha) \approx 1 - \Phi(\sqrt{n}\alpha, 1)$$
$$= \frac{1}{\sqrt{\pi}} \int_{\sqrt{n}\alpha}^{\infty} e^{-t^2/2} dt$$

The integral can be approximated by an asymptotic series through integration by parts[a]:

$$\int_x^\infty e^{-t^2} dt = -\frac{1}{2} \left(\frac{e^{-t^2}}{t} \Big|_x^\infty + \int_x^\infty \frac{e^{-t^2}}{t^2} dt \right)$$
$$= e^{-x^2} \left(\frac{1}{2x} + \mathcal{O}(x^{-3}) \right)$$

yielding an exponential tail with rate $\alpha^2/2$ in accordance with the result obtained in a).

[a] For more details on this and other approximations of the error function see for example http://mathworld.wolfram.com/Erf.html

12 Sampling bias

20 pts.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample of iid. random variables. Each X_i is distributed according to Q, i.e

$$\mathrm{P}(X_j = x) = Q(x)$$

In the limit of $n \to \infty$ we would find that the empirical distribution $P_{\mathbf{X}} \to Q$.

What if we skew the sampling process by constraining it to a some event $E \subset \mathcal{P}$? If the sample does not fulfill E it is discarded. What is the distribution of X_i then? Thus we are interested in the conditional probability

$$\mathrm{P}(X_1 = x | P_{\mathbf{X}} \in E). \tag{3}$$

The (or rather a) *conditional limit theorem*[1] assures that the sought after distribution is again given by

$$P^* = \arg\min_{P \in E} D(P||Q) \tag{4}$$

in the limit of large n, i.e

$$\mathrm{P}(X_1 = x | P_{\mathbf{X}} \in E) \to P^*(x) \text{ (in probability)}.$$

As an example consider $X_i \sim \mathcal{U}([0, 1])$ uniformly distributed on the interval $[0, 1]$[2].

For whatever reason we decide to only include those $\mathbf{X}$ with $\langle X \rangle \geq \alpha$ and $\mathrm{Var}(\mathbf{X}) \geq \beta^2$.

(a) *20pt* - Show that

$$p^*(x) = \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2\right)$$

by again maximizing the KL-divergence and find $\lambda_{0,1,2}$ as functions of α, β!

(b) *Bonus 0pt* - It is instructive and very easy to simulate such a sampling process on the computer for different sets of parameters. Implement a routine in your favorite language and plot the distribution in a histogram!

One set of parameters that works well is

$$\alpha = 0.5, \quad \beta = 0.1, \quad n \approx 10$$

You probably want to draw in the order of 10^6 samples to generate a histogram.

It is interesting to see that the limiting distribution arises already for quite small values of n. You can also easily explore the behavior for non-uniform Q.

[1] For the exact statement and prove see *Cover & Thomas* Ch. 11

[2] You might be worried that we are suddenly dealing with continuous variables. Just as with the differential entropy, one obtains analogous statements by quantizing and taking an appropriate limit.

Solution

An implementation of the sampling is available from the website. If you do not have or do not want a local installation of Julia (available from http://www.julialang.org), you can run the notebook at http://juliabox.com.

(a) The calculation to obtain the form of the distribution is the same as in the previous problem. Unfortunately it is not possible to obtain expressions for the parameters in closed form. The constraints can be evaluated, but the resulting expressions involve various error-functions and are not solvable explicitly. One needs to resort to numerical methods. Conceptually one looks for a set of parameters λ_1, λ_2 such that

$$\alpha = \int_0^1 xp(x;\lambda_1,\lambda_2)dx \qquad \beta = \int_0^1 x^2 p(x;\lambda_1,\lambda_2)dx - \alpha^2$$

where

$$p(x;\lambda_1,\lambda_2) = \frac{\exp(\lambda_1 x + \lambda_2 x^2)}{Z(\lambda_1,\lambda_2)} \tag{5}$$

with partition function

$$Z(\lambda_1,\lambda_2) = \int_0^1 \exp(\lambda_1 x + \lambda_2 x^2)dx$$

In the notebook (link on the website) you find outlined two routes to determine the parameters. First, one may sample a histogram under the constraints and fit it to the prescribed exponential form. Note that this is an approximation, but for large values of n we know that the distribution converges to the form given by (5). This brings its own problem, because for large n, the events become exponentially rarer and a *huge* number of samples is required.

The better approach is to numerically solve the system of non-linear equations, which should yield the exact solution. The notebook compares both routes.

Note: When you play around with the parameters α, β you might come up with strange numerical solutions. In particular it might happen, that the parameters have the wrong sign. If this is the case, check whether the constraints are already (partially) fulfilled! For example, the underlying uniform distribution Q has variance $1/3 - 1/4 \approx 0.08$. When the constraint is below that value, one must drop the constraint altogether to get a sensible answer. Look at the above figure showing the probability simplex. The whole minimization procedure relies on finding P^* at the boundary, which in turn implies that Q must not fulfill any inequality constraint on its own.

The method of Lagrange multipliers applies as is only to equality constraints. In problem eleven we argued that the minima are necessarily found for tight constraints because Q lies outside the set of constraints.

There exist generalizations of Lagrange's method. In particular the *Karush-Kuhn-Tucker conditions*[a] apply to inequality constraints. Pay special attention to what is called *complementary slackness*. In words it means that either the constraints are tight or the multiplier must be zero.

(b) *see notebook and (a)*

[a] https://en.wikipedia.org/wiki/KarushKuhnTucker_conditions