Lecture-12: The Gärtner-Ellis Theorem

1 The Gärtner-Ellis Theorem

This theorem is a powerful result which establishes the existence of a large deviation principle for processes where the cumulant generating function tend towards a well behaved limit implying not-too-strong dependence between successive values. It has several formulations out of that we will state a simplified version of the general theorem which not required very high technical definitions.

Definition 1.1. For any function $F : \mathbb{R} \to \mathbb{R}$, we say that $x \in \mathbb{R}$ is an **exposed point** of the function F if there exists $t \in \mathbb{R}$ such that $ty - F(y) > tx - F(x)$ for any $y \neq x$.

If F is **convex**, a sufficient condition for x to be an exposed point is that F is twice differentiable at x, with $F''(x) > 0$.

Theorem 1.2 (Gärtner-Ellis). *Consider a function $f : \mathcal{X}^N \to \mathbb{R}$. We assume that for an N-length random sequence $X \in \mathcal{X}^N$ the normalized log moment generating function $\psi_N(t) = \frac{1}{N} \log \mathbb{E} e^{tNf}$ exists, and has a finite limit $\psi(t) = \lim_{N\to\infty} \psi_N(t)$, for any $t \in \mathbb{R}$. Let I_ψ be the inverse Legendre transform and $\mathcal{E}$ be the corresponding set of exposed points of the function I_ψ, then the following hold.*

1. *For any closed set $F \in \mathbb{R}$,*
$$\limsup_{N\to\infty} \frac{1}{N} \log P_N \{f \in F\} \leqslant - \inf_{f\in F} I_\psi(f).$$
2. *For any open set $G \in \mathbb{R}$,*
$$\limsup_{N\to\infty} \frac{1}{N} \log P_N \{f \in G\} \geqslant - \inf_{f\in G\cap\mathcal{E}} I_\psi(f).$$
3. *If $\psi(t)$ is differentiable for any $t \in \mathbb{R}$, then the last statement holds true if the* inf *is taken over the whole set G (rather than over $G \cap \mathcal{E}$).*

Proof. Since the Legendre transform is convex in t and its inverse is convex in f, we can write the Legendre transform as the Legendre transform of its inverse.

Involutive Property of Legendre Transform

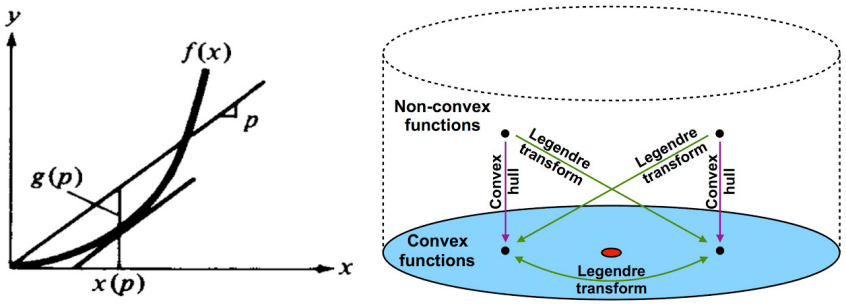
Arnold gives a geometric proof based on the fact that if $g(p) = \mathcal{L}(f)(p)$ then the graph $y = xp - g(p)$ is the tangent to the slope p to the graph $y = f(x)$. Since $f(x)$ is convex, all the tangent lines are below the graph, so if we fix $x = x_0$, the maximal value of $x_0 p - g(p)$ as a function of p is $f(x_0)$ (otherwise we are below the graph, instead of on it). Thus $\mathcal{L}(g)(x_0) = \sup_p(x_0 p - g(p)) = f(x_0)$

We can prove this algebraically. Noe let $g(p) = \mathcal{L}(f)(p)$ and we compute $\mathcal{L}(\mathcal{L}(f))(x) = \mathcal{L}(g)(x)$. $\mathcal{L}(g)(x) = \sup_{p(x)}(x.p(x) - g(p(x)))$, where $p(x)$ is defined by $x = (\nabla g)(p(x))$. Now, g is defined by $g(p) = \sup_{y(p)}(p.y(p) - f(y(p)))$, where $y(p)$ is defined by $p = (\nabla f)(y(p))$. Therefore, we have

$$x = (\nabla g)(p(x)) = y(p(x))$$

and

$$\begin{aligned}\mathcal{L}(g)(x) &= \sup_{p(x)}(x.p(x) - g(p(x))) \\ &= y(p(x)).p(x) - [p(x).y(p(x)) - f(y(p(x)))] \\ &= f(x)\end{aligned}$$



That is,

$$I_\psi(f) = \sup_{t\in\mathbb{R}}[tf - \psi(t)], \qquad \psi(t) = \sup_f[tf - I_\psi(f)].$$

Since $\psi(t) = \lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}e^{tNf}$, it follows that the large-deviation principle holds for the function f with a rate function $I_\psi(f)$, such that $P_N(f) \doteq e^{-NI_\psi(f)}$. □

The inverse Legendre transform yields an upper bound on the probability of a large fluctuation of the macroscopic observable. This upper bound is tight unless a 'first-order phase transition' occurs, corresponding to a discontinuity in the first derivative of $\psi(t)$, as we saw in the low-temperature phase of the Curie-Weiss model. It is worth mentioning that $\psi(t)$ can be non-analytic at a point t^* even though its first derivative is continuous at t^*. This corresponds, to a 'higher-order' phase transition.

1.1 Typical sequences

From the previous lectures we know clearly the concept of typical sequences, more precisely, we want to investigate the large deviation of the probability itself. We can re-write the log moment generation function for the empirical entropy $r(x) = -\frac{1}{N}\log P_N(x)$, as

$$\begin{aligned}\psi_N(t) &= \frac{1}{N}\log\mathbb{E}[e^{tNf}] = \frac{1}{N}\log\sum_x P_N(x)e^{tNr(x)}\\ &= \frac{1}{N}\log\sum_x P_N(x).P_N(x)^{-t} = \frac{1}{N}\log\sum_x P_N(x)^{1-t}.\end{aligned}$$

Let $P_N(x) = \frac{1}{Z_N(\beta)}e^{-\beta E_N(x)}$ be the Boltzmann distribution with the energy function $E_N(\underline{x})$ and the partition function $Z_N(\beta) = \sum_x e^{-\beta E(\underline{x})}$, then in terms of the free-energy density $f_N(\beta) = -\frac{1}{N}\log Z_N(\beta)$, we can write

$$\begin{aligned}\psi_N(t) &= \frac{1}{N}\log\sum_x\left[\frac{e^{-\beta E_N(x)}}{Z_N(\beta)}\right]^{1-t}\\ &= \frac{1}{N}(\log\sum_x e^{-\beta(1-t)E_N(x)} - (1-t)\log Z_N(\beta))\\ &= \beta(1-t)f_N(\beta) - \beta f_N((1-t)\beta).\end{aligned}$$

Assuming that the thermodynamic limit $f(\beta) = \lim_{N\to\infty} f_N(\beta)$ exists and is finite, It follows that the Legendre transform $\psi(t)$ exists for the empirical entropy. We can apply the Gärtner-Ellis theorem to compute the probability of a large fluctuation of the empirical entropy $r(\underline{x})$.

$$\psi(t) == \beta(1-t)f(\beta) - \beta f((1-t)\beta).$$

As long as $f(\beta)$ is analytic, large fluctuations are exponentially rare and the asymptotic equipartition property of independent random variables is essentially recovered. This follows from the fact that $\mathbb{E}_P[r(X)] = h(P_N) = \frac{1}{N}H(P_N)$, and the set of ϵ-typical sequences is

$$T_{N,\epsilon} = \left\{x \in \mathcal{X}^N : |r(X) - h(P_N)| \leqslant \epsilon\right\}.$$

1. Under certain conditions, we can show that $\lim_{N\to\infty} P_N\{X \in T_{N,\epsilon}\} = 1$.

2. From the definition of $r(\underline{x})$ and $T_{N,\epsilon}$, it follows that for any $x \in T_{N,\epsilon}$

$$2^{-N(h(P_N)+\epsilon)} \leqslant P_N(\underline{x}) \leqslant 2^{-N(h(P_N)-\epsilon)}.$$

On the other hand, if there is a phase transition at $\beta = \beta_c$, where the first derivative of $f(\beta)$ is discontinuous, then the likelihood $r(x)$ may take several distinct values with a non-vanishing probability. The same thing can be seen from Curie - Weiss Model.

Example 1.3. consider a **Markov Chain** $X_0, X_1, ..., X_i, ...$ taking values in a finite state space $\mathcal{X}$, and assume all the elements of transition matrix $w(x \to y)$ to be strictly positive. Compute the large deviation properties of of the empirical average $\frac{1}{N}\sum_i f(X_i)$. One can show that the limit moment generating function $\psi_(t)$, exists, and can be computed using the following recipe. Define the 'tilted' transition probabilities as $w_t(x \to y) = w(x \to y)\exp[tf(y)]$. Let $\lambda(t)$ be the largest solution of the eigenvalue problem.

$$\begin{aligned}
\psi(t) &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}[\exp[tNf]] \\
&= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}[\exp[t\sum_{i=1}^{N} f(x_i)]] \\
&= \lim_{N\to\infty} \frac{1}{N} \log \sum_{X\in\mathcal{X}^N} \exp[t\sum_{i=1}^{N} f(x_i)] P_{x_0} \prod_{i=1}^{N} w(x_{i-1} \to x_i) \\
&= \lim_{N\to\infty} \frac{1}{N} \log \sum_{x\in\mathcal{X}^N} P_{x_0} \prod_{i=1}^{N} w(x_{i-1} \to x_i) \exp[tf(x_i)]
\end{aligned}$$

This can be reduced to matrix multiplication.

$$\psi(t) = \lim_{N\to\infty} \frac{1}{N} \log P_0 W_t^N 1$$

The moment generating function is simply given by $\psi(t) = \log\lambda(t)$. This result came from the Perron - Frobenius Theorem.

The result of Perron-Frobenius theorem is convergence to steady state of homogeneous markov chain is geometric with relative speed equal to the magnitude of the largest eigen value. The same result can be found by applying SVD in the expression for $\psi(t)$.

2 The Gibbs Free Energy

We provide a motivation for the Boltzmann distribution to be a natural choice for probability distribution of the configuration of a physical system.

2.1 Variational principle

Consider a system with a configuration space $\mathcal{X}$, and an energy function $E : \mathcal{X} \to \mathbb{R}$. The Boltzmann distribution is

$$\mu_\beta(x) = \frac{e^{-\beta E(x)}}{Z(\beta)} = \exp\left(-\beta(E(x) + \frac{1}{\beta}\log Z(\beta))\right) = e^{-\beta(E(x)-F(\beta))},$$

where the 'free energy' $F(\beta)$, is a function of the inverse temperature β defined by the fact that $\sum_{x\in\mathcal{X}} \mu_\beta(x) = 1$.

Definition 2.1. We define the **Gibbs free energy** $G : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ as the following real-valued functional over the space of probability distributions on $\mathcal{X}$

$$G[P] = \sum_{x \in \mathcal{X}} P(x)E(x) + \frac{1}{\beta} \sum_{x \in \mathcal{X}} P(x) \log P(x).$$

The Gibbs free energy should not be confused with the free energy $F(\beta)$.

Proposition 2.2. *The Gibbs free energy $G : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ is a convex functional, and it achieves its unique minimum on the Boltzmann distribution $P = \mu_\beta$. Moreover, $G[\mu_\beta] = F(\beta)$, where $F(\beta)$ is the free energy.*

Proof. It is easy to rewrite the Gibbs free energy in terms of the KL divergence between P and the Boltzmann distribution μ_β

$$G[P] = \frac{1}{\beta} \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{e^{-\beta E(x)}} = \frac{1}{\beta} D(P \| \mu_\beta) + F(\beta).$$

□

The relation between the Gibbs free energy and the KL divergence implies a simple probabilistic interpretation of the Gibbs variational principle. Imagine that a large number $\mathcal{N}$ of copies of the same physical system have been prepared. Each copy is described by the same energy function $E(x)$. Now consider the empirical distribution P of the $\mathcal{N}$ copies. Typically, P will be close to the Boltzmann distribution μ_β. Sanov's theorem implies that the probability of an 'atypical' distribution is exponentially small in $\mathcal{N}$:

$$\mathbb{P}[P] \doteq \exp\left(-\mathcal{N}(G[P] - F(\beta))\right).$$

When the partition function of a system cannot be computed exactly, the above result suggests a general line of approach for estimating the free energy: one can minimize the Gibbs free energy in some restricted subspace of 'trial probability distributions' P. These trial distributions should be simple enough that $G[P]$ can be computed, but the restricted subspace should also contain distributions which are able to give a good approximation to the true behavior of the physical system. For each new physical system one will thus need to find a good restricted subspace.

2.2 Mean-field approximation

Mean-field approximation is taking the class of distributions over independent variables as the trial family.

Example 2.3 (Ising Model). Consider particles on the lattice $\mathbb{L}$ of nodes $[L]^d$ and edges $((i,j) : |i-j| = 1)$. Each particle at node i has spin $\sigma_i \in \mathcal{X} = \{-1, 1\}$. The energy function under the external magnetic field B is given by

$$E(\sigma) = -\frac{1}{d} \sum_{(i,j)} \sigma_i \sigma_j - B \sum_i \sigma_i.$$

We assume periodic boundary conditions, and choose the trial family of distributions to be

$$Q_m(\sigma) = \prod_i q_m(\sigma_i),$$

where $q_m(\sigma_i) = \frac{(1+m)}{2} \mathbb{1}_{\{\sigma_i = 1\}} + \frac{(1-m)}{2} \mathbb{1}_{\{\sigma_i = -1\}}$ for some $m \in [-1, 1]$. That is, under the distribution Q_m, the spins are i.i.d. with mean m.

We can find the density of Gibbs free energy as

$$g(m; \beta, B) \triangleq \frac{G[Q_m]}{|L|^d} = -\frac{1}{2} m^2 - Bm - \frac{1}{\beta} \mathcal{H}\left(\frac{1+m}{2}\right).$$

From the Gibbs variational principle, we have

$$f_d(\beta, B) \leqslant \inf_m g(m; \beta, B) = f_{\mathrm{CW}}(\beta, h) - \frac{1}{2}.$$

Indeed, the mean-field approximation becomes better the larger the dimension d, and it is asymptotically exact for $d \to \infty$.