Sanov's theorem for sub-sampling from individual sequences

Amir Dembo
Department of Mathematics and Department of Statistics
Stanford University

Ofer Zeitouni
Department of Electrical Engineering
Technion- Israel Institute of Technology, Haifa, ISRAEL

September 8, 1994

1 Main Result

Let $x = (x_1, x_2, \ldots, x_m, \ldots)$ be a $\Sigma$-valued deterministic sequence such that $L_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ converge to $P_x$ weakly. Consider the following random sub-sampling scheme. Fix $\beta \in (0, 1)$, and $m = m(n)$ such that $n/m \to \beta$, generating the random variables $X_1^n, \ldots, X_n^n$ by sampling $n$ values out of $(x_1, \ldots, x_m)$ without replacement, i.e. $X_i^n = x_{j_i}$ for $i = 1, \ldots, n$ where each choice of $j_1 \neq j_2 \neq \cdots \neq j_n \in \{1, \ldots, m\}$ is equally likely (and independent of the sequence $x$).

The next proposition shows that perhaps somewhat surprisingly (see Remark 1 immediately following its statement), the large deviations of the empirical measure of the resulting sample admits a rate function which is independent of the particular sequence $x$ but different from the

*Partially supported by NSF DMS92-09712 grant and by a US-ISRAEL BSF grant.
†Partially supported by a US-Israel BSF grant and by the Army Research Office under grant ARO DAAL03-92-G-0115 (Center for Intelligent Control Systems).

Key Words: Large deviations, sampling without replacement, empirical measure.

AMS (1991) subject classifications: 60F10
rate function of Sanov's theorem.

**Proposition 1.1** The sequence $L_m^n = n^{-1} \sum_{i=1}^m \delta_{x_i}$ satisfies the large deviation principle (LDP) in $M_1(\Sigma)$ with the convex good rate function

$$I(\nu\beta, P_X) = \begin{cases} H(\nu|P_X) + \frac{1-\beta}{\beta} H \left( \frac{P_X - \nu}{1-\beta} \right) |P_X) & \text{if } P_X - \frac{\nu}{1-\beta} \in M_1(\Sigma) \\ \infty & \text{otherwise}. \end{cases}$$

**Remarks:**

1) Consider the probability space $(\Omega_1 \times \Omega_2, \mathcal{B} \times \mathcal{B}_\Sigma, P_1 \times P_2)$, with $\Omega_2 = \Sigma^\mathbb{N}$, $P_2$ stationary and ergodic with marginal $P_X$ on $\Sigma$, and let $\omega_2 = x = (x_1, x_2, \ldots, x_m, \ldots)$ be a realization of an infinite sequence under the measure $P_2$. $(\Omega_1, \mathcal{B}, P_1)$ represents the randomness involved in the sub-sampling. Since $\Sigma$ is Polish, by the ergodic theorem the empirical measures $L_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ converge to $P_X$ weakly for $(P_2)$ almost every $\omega_2$. It follows that Proposition 1.1 may be applied for almost every $\omega_2$, yielding the same LDP for $L_m^n$ under the law $P_1$ for almost every $\omega_2$. Note that for $P_2$ a product measure (corresponding to an i.i.d. sequence), the LDP for $L_m^n$ under the law $P_1 \times P_2$ is given by Sanov's theorem and admits a different rate function.

2) Using a projective limit approach, the LDP for the empirical measures in sampling without replacement is derived in [2, Section 7.2] assuming that $L_m \rightarrow P_X$ in the $\tau$-topology. In the context of sub-sampling described in the previous remark this assumption fails as soon as $P_X$ is non-atomic, and a completely different method of proof is thus needed.

Let $g_{\beta}(x) = (1 - \beta x)/(1 - \beta)$ and denote by $M_+^\Sigma$ the space of all non-negative, finite Borel measures on $\Sigma$. The first step in the proof of Proposition 1.1 is to derive the LDP for a closely related sequence of empirical measures of deterministic positions and random weights which is much simpler to handle.

**Lemma 1.3** Let $J_i$ be i.i.d. Bernoulli(\(\beta\)) random variables, and $x_i \in \Sigma$ non-random such that $m^{-1} \sum_{i=1}^m \delta_{x_i} \rightarrow P_X$ weakly in $M_1(\Sigma)$. Then, the sequence $L'_n = n^{-1} \sum_{i=1}^m J_i \delta_{x_i}$ satisfies the LDP in $M_+^\Sigma$, equipped with the weak ($C_0(\Sigma)$-)topology, with the convex good rate function

$$I(\nu) = \begin{cases} \int f \log f dP_X + \frac{1-\beta}{\beta} \int g_{\beta}(f) \log g_{\beta}(f) dP_X & \text{if } \nu \in M_+^\Sigma(\Sigma) \text{ and } f = \frac{\nu}{P_X} \leq \frac{1}{\beta} \\ \infty & \text{otherwise}. \end{cases}$$

2
Proof: Clearly, \( I(\cdot) \geq 0 \) is convex, and the convex set \( \Psi = \{ f \in L_1(P_X) : I(\nu) \leq \alpha, f = \frac{d\nu}{dP_X} \} \) is uniformly integrable, hence weakly sequentially compact as a subset of \( L_1(P_X) \). Note that \( F_\beta = \{ f \in L_1(\mu) : 0 \leq f \leq \frac{1}{\beta} \} \) is a closed set, and \( f \mapsto g_\beta(f) \) a continuous mapping between \( F_\beta \) and \( F_{1-\beta} \). Since \( f \mapsto \int f \log f dP_X : F_\beta \to \mathbb{R} \) is continuous for every fixed \( b < \infty \), it follows that \( \Psi \) is closed in \( L_1(P_X) \), and by convexity it is also weakly closed and hence weakly compact in \( L_1(P_X) \).

Since weak convergence in \( L_1(P_X) \) gives rise to convergence in \( M_+(\Sigma) \) of the associated measures, it follows that \( \{ \nu : I(\nu) \leq \alpha \} \) is compact.

For \( \phi \in C_b(\Sigma) \) we have
\[
\log E[\exp(n \int \phi dL'_n)] = \log E[\exp(\sum_{i=1}^{m} \phi_i(x_i))] = \sum_{i=1}^{m} \log(\beta e^{\phi(x_i)} + 1 - \beta),
\]
implying that
\[
\Lambda(\phi) \triangleq \lim_{n \to \infty} \frac{1}{n} \log E[\exp(n \int \phi dL'_n)] = \frac{1}{\beta} \int \log(\beta e^{\phi(x)} + 1 - \beta) P_X(dx).
\]
Let \( \mathcal{X} \) be the algebraic dual of \( C_b(\Sigma) \) equipped with the \( C_b(\Sigma) \)-topology, and for \( \vartheta \in \mathcal{X} \) define
\[
\Lambda^*(\vartheta) = \sup_{\phi \in C_b(\Sigma)} \{ \langle \phi, \vartheta \rangle - \Lambda(\phi) \}.
\]
Consider the \( \mathbb{R}^k \)-valued random variables \( \hat{S}_n = (\int \phi_1 dL'_n, \ldots, \int \phi_k dL'_n) \) for fixed \( \phi_1, \ldots, \phi_k \in C_b(\Sigma) \) and observe that they have the limiting logarithmic moment generating function
\[
\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E[\exp(n(\lambda, \hat{S}_n))] = \frac{1}{\beta} \int \log(\beta e^{\sum_{i=1}^{k} \lambda_i \phi_i(x)} + 1 - \beta) P_X(dx).
\]
The function \( \Lambda(\lambda) \) is finite and differentiable in \( \lambda \) throughout \( \mathbb{R}^k \) for any collection \( \phi_1, \ldots, \phi_k \in C_b(\Sigma) \). Hence, by [2, Corollary 4.6.11 part (a)], the sequence \( L'_n \) satisfies the LDP in \( \mathcal{X} \) with the good rate function \( \Lambda^*(\cdot) \).

Identify \( M_+(\Sigma) \) as a subset of \( \mathcal{X} \). Fix \( \nu \in M_+(\Sigma) \) such that \( f = \frac{d\nu}{dP_X} \leq \frac{1}{\beta} \), and observe that for every \( \phi \in C_b(\Sigma) \)
\[
\int \phi d\nu - I(\nu) = \Lambda(\phi) - \frac{1}{\beta} \int h \left( 1 - \beta f \frac{1 - \beta}{\beta e^{\phi} + 1 - \beta} \right) dP_X, \tag{1.5}
\]
where \( h(x|p) = x \log(x/p) + (1 - x) \log((1 - x)/(1 - p)) \) for \( x, p \in [0, 1] \). Since \( h(x|p) \geq 0 \) with equality iff \( x = p \), it follows by the choice \( f = \frac{e^{\phi}}{e^{\phi} + 1 - \beta} \) in (1.5) that
\[
\Lambda(\phi) = \sup_{\nu \in M_+(\Sigma)} \{ \int \phi d\nu - I(\nu) \} = \sup_{\vartheta \in \mathcal{X}} \{ \langle \phi, \vartheta \rangle - I(\vartheta) \},
\]
implying by duality that $I(\cdot) = \Lambda^*(\cdot)$ (see [2, Lemma 4.5.8]). In particular, $L'_n$ thus satisfies the LDP in $\mathcal{M}_+(\Sigma)$ (see [2, Lemma 4.1.5 part (b)]) with the convex good rate function $I(\cdot)$.

**Proof of Proposition 1.1:** Note that for every $\nu \in \mathcal{M}_1(\Sigma)$, $I(\nu)$ of (1.4) equals to $I(\nu|\beta, P_X)$ of (1.2) which is thus a convex good rate function. Use $(x_1, x_2, \ldots)$ to generate the sequence $L'_n$ as in Lemma 1.3. Let $V_n$ denote the number of i-s such that $J_i = 1$, i.e. $V_n = nL'_n(\Sigma)$. The key to the proof is the following coupling. If $V_n > n$ choose (by sampling without replacement) a random subset $\{i_1, \ldots, i_{V_n - n}\}$ among those indices with $J_i = 1$ and set $J_i$ to zero on this subset. Similarly, if $V_n < n$ choose a random subset $\{i_1, \ldots, i_{n - V_n}\}$ among those indices with $J_i = 0$ and set $J_i$ to one on this subset. Re-evaluate $L'_n$ using the modified $J_i$ values and denote the resulting (random) probability measure by $Z_n$. Note that $Z_n$ has the same law as $L''_n$ which is also the law of $L'_n$ conditioned on the event $\{V_n = n\}$. Since $V_n$ is a Binomial$(m, \beta)$ random variable, and $n/m \to \beta \in (0, 1)$ it follows that

$$\liminf_{n \to \infty} \frac{1}{n} \log P(V_n = n) = 0.$$  

Fix a closed set $F \subset \mathcal{M}_1(\Sigma)$ and observe that $P(L''_n \in F) = P(Z_n \in F) \leq P(L'_n \in F)/P(V_n = n)$ implying that $\{L''_n\}$ satisfies the large deviations upper bound in $\mathcal{M}_1(\Sigma)$ with the rate function $I(\cdot|\beta, P_X)$. Let $\mathcal{F}_{LU}$ denote the class of Lipschitz continuous functions $f : \Sigma \to \mathbb{R}$, with Lipschitz constant and uniform bound 1. Recall that $\beta(P, Q) = \sup_{f \in \mathcal{F}_{LU}} |\int f dP - \int f dQ|$ is a metric on $\mathcal{M}_+(\Sigma)$ which is equivalent to the $C_b(\Sigma)$-topology (for a proof of this elementary fact, see [1, Lemma 6]). Since for $\nu \in \mathcal{M}_1(\Sigma)$

$$\beta(Z_n, L'_n) = n^{-1}|V_n - n| = |L'_n(\Sigma) - 1| \leq \beta(L'_n, \nu),$$

it follows that

$$P(\beta(L''_n, \nu) < 2\delta) = P(\beta(Z_n, \nu) < 2\delta) \geq P(\beta(L'_n, \nu) < \delta, \beta(Z_n, L'_n) < \delta) = P(\beta(L'_n, \nu) < \delta).$$

Consequently, by the LDP of Lemma 1.3, for every $\nu \in \mathcal{M}_1(\Sigma)$ and all $\delta > 0$

$$\liminf_{n \to \infty} \frac{1}{n} \log P(\beta(L''_n, \nu) < \delta) \geq -I(\nu) = -I(\nu|\beta, P_X).$$

This completes the proof of the large deviations lower bound (since the metric $\beta(\cdot, \cdot)$ is also equivalent to the weak topology on $\mathcal{M}_1(\Sigma)$).
References
