Model-fitting in the presence of outliers

Jayakrishnan Unnikrishnan
Audiovisual Communications Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
Email: jay.unnikrishnan@epfl.ch

***Abstract*—We study the problem of parametric model-fitting in a finite alphabet setting. We characterize the weak convergence of the goodness-of-fit statistic with respect to an exponential family when the observations are drawn from some alternate distribution. We then study the effects of outliers on the model-fitting procedure by specializing our results to ϵ-contaminated versions of distributions from the exponential family. We characterize the sensitivity of various distributions from the exponential family to outliers, and provide guidelines for choosing thresholds for a goodness-of-fit test that is robust to outliers in the data.**

I. Introduction

Consider a sequence of random observations $\boldsymbol{Z} = \{Z_i : i = 1, \ldots, n\}$ drawn i.i.d. from a finite alphabet $\mathsf{Z} = \{z_1, , z_2, \ldots, z_N\}$. Let $\mathcal{P}(\mathsf{Z})$ denote the space of probability measures on Z. The problem of parametric model-fitting is concerned with the following question: what is the most likely distribution from a parametric class of probability distributions

$$\{\pi_\theta : \theta \in \mathbb{R}^d\} \subset \mathcal{P}(\mathsf{Z})$$

that could have generated the string $\boldsymbol{Z}$? The commonly used solution to this problem is given by the maximum-likelihood (ML) estimate $\hat{\theta}_n$ of the parameter based on the observations.

A related problem that arises in model-fitting is to quantify the accuracy of the fit, often called the goodness-of-fit of the model. This question is typically answered using the limiting distributions of some distance metric between the empirical distribution of the observations and the ML distribution. For instance, in the finite alphabet setting, a useful metric is the Kullback-Leibler divergence $D(\Gamma^n \| \pi_{\hat{\theta}_n})$ where Γ^n denotes the empirical distribution (type) of the observations:

$$\Gamma^n(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i = z\}, \qquad z \in \mathsf{Z} \tag{1}$$

where $\mathbb{I}$ is the indicator function. The idea is to accept the null hypothesis that the observations were indeed drawn from some distribution in the parametric family $\{\pi_\theta\}$ only if the divergence $D(\Gamma^n \| \pi_{\hat{\theta}_n})$ is less than some pre-decided threshold τ chosen to meet some false alarm probability constraint. The threshold τ is typically chosen based on the following asymptotic weak convergence of the test statistic under the null hypothesis

$$2nD(\Gamma^n \| \pi_{\hat{\theta}_n}) \xrightarrow[n\to\infty]{d.} \chi^2_{N-d-1}$$

where χ^2_a denotes a chi-square random variable with a degrees of freedom and N is the alphabet size. This result, which holds under some regularity conditions on $\{\pi_\theta\}$, enables us to obtain approximations to the distribution of the test statistic for large n which can then be used to set thresholds τ that approximately meet a target false alarm constraint.

One of the pitfalls in using the above technique to quantify goodness-of-fit is its sensitivity to model inaccuracies. In reality all models are only approximate. One of the common irregularities is the presence of outliers in the data. In this paper, we model data with outliers as coming from *ϵ-contamination classes* of distributions [1] from an exponential family and study the behavior of the test statistic $D(\Gamma^n \| \pi_{\hat{\theta}_n})$ in the presence of outliers. We first obtain the asymptotic distribution of the divergence statistic when the observations are drawn from an arbitrary distribution outside of the exponential family. We then specialize our results to the case of distributions from the ϵ-contamination class and characterize the behavior of the test statistic as a function of ϵ. These results quantify the sensitivity of the test to outliers and provide guidelines on how to choose the threshold τ to design a test that is robust to outliers. Our results are also useful for approximating the power of the goodness-of-fit test to reject general distributions that lie outside the exponential family.

Most of the probabilistic studies of outliers in model-fitting have focussed on robust estimation, regression and hypothesis testing [1], [2], [3]. In this paper we pursue a detailed study of the effect of outliers on goodness-of-fit testing. A related work is the robust goodness-of-fit test that we studied in [4]. The finite alphabet uncertainty model studied in [4] is however not suited for analyzing the effects of data outliers.

We describe the problem setup in Section II, our results in Section III and conclusions in Section IV. Throughout the paper, we use the following notation: For measures $\mu \in \mathcal{P}(\mathsf{Z})$ we use $\mu(z)$ to denote the mass at $z \in \mathsf{Z}$. We sometimes use μ also to denote the vector in $\mathbb{R}^N$ with $\mu_i = \mu(z_i)$. For functions f defined on Z we denote the expected value under μ by $\langle \mu, f \rangle := \sum_{z \in \mathsf{Z}} \mu(z) f(z)$.

II. Problem Setup

Consider the following exponential family of distributions.

$$\begin{aligned} \mathcal{E}_\pi &:= \{\pi_\theta : \theta \in \mathbb{R}^d\}, \\ \text{where } \pi_\theta(z) &:= \pi(z) \exp(\theta^T \psi(z) - \Lambda(\theta)),\ z \in \mathsf{Z} \end{aligned} \tag{2}$$

where $\pi \in \mathcal{P}(\mathsf{Z})$ has full support on Z and

$$\psi := (\psi_1, \psi_2, \ldots, \psi_d)^T$$

is a vector of affinely independent real-valued functions over Z, i.e. the functions $\{\psi_0, \psi_1, \psi_2, \ldots, \psi_d\}$ are linearly independent over Z where $\psi_0(z) = 1$ for all $z \in \mathsf{Z}$. Hence $\theta^T\psi = \sum_{i=1}^d \theta_i\psi_i$ is a real-valued function on Z. The function $\Lambda(.)$ in (2) is defined by

$$\Lambda(\theta) = \log\langle \pi, \exp(\theta^T\psi)\rangle.$$

Clearly Λ is an analytic function over $\mathbb{R}^d$. It is also easy to see that

$$\langle \pi_\theta, \psi\rangle = \nabla\Lambda(\theta). \tag{3}$$

Furthermore, the Hessian $\nabla^2\Lambda$ is given by the covariance matrix of the random vector $\psi(Z)$ when $Z \sim \pi_\theta$:

$$\nabla^2\Lambda(\theta) = \mathrm{Cov}_{\pi_\theta}\psi(Z) \tag{4}$$

and is positive definite everywhere on $\mathbb{R}^d$ under the affine independence of the functions in ψ (see [5, Lemma III.1]).

Let $\hat{\theta}_n$ denote the ML estimate of the parameter θ based on the n i.i.d. observations $\{Z_1, Z_2, \ldots, Z_n\}$. In this paper we are interested in testing whether or not these observations were drawn from some distribution in $\mathcal{E}_\pi$; i.e., we are testing the following composite null hypothesis

$$\mathcal{H}_0 : Z_i \sim \text{i.i.d.}\, \mu, \quad i = 1, 2, \ldots, \text{ for some } \mu \in \mathcal{E}_\pi. \tag{5}$$

The test statistic typically used for this purpose is the divergence $D(\Gamma^n\|\pi_{\hat{\theta}_n})$ between the empirical distribution and the ML distribution. We accept hypothesis $\mathcal{H}_0$ if this statistic is below some fixed threshold. i.e., the test is of the form

$$\hat{\mathcal{H}} = \mathbb{I}\{D(\Gamma^n\|\pi_{\hat{\theta}_n}) < \tau\} \tag{6}$$

with $\hat{\mathcal{H}} = 0$ indicating a decision in favor of $\mathcal{H}_0$. This test statistic can be motivated based on its interpretation in terms of error exponents as we elaborate in Section III-C. In the rest of this paper we study the asymptotic behavior of the statistic $D(\Gamma^n\|\pi_{\hat{\theta}_n})$ and its implications for goodness-of-fit testing in the presence of outliers.

III. Asymptotic analysis

Before we proceed, we need the following known result.

Lemma III.1. *Let $\mu \in \mathcal{P}(\mathsf{Z})$ be any distribution with full support over Z and π_θ be as defined in (2). Then there exists $\theta^* \in \mathbb{R}^d$ that solves the following reverse I-projection problem*

$$\inf_{\theta\in\mathbb{R}^d} D(\mu\|\pi_\theta). \tag{7}$$

Proof: Let $\hat{\pi}$ denote the I-projection

$$\hat{\pi} = \underset{\nu\in\mathcal{P}(\mathsf{Z}):\langle\nu,\psi\rangle=\langle\mu,\psi\rangle}{\arg\min} D(\nu\|\pi)$$

Clearly, the minimizer $\hat{\pi}$ exists since we are optimizing over a compact set. Furthermore, by the Lagrange multiplier theorem it follows that $\hat{\pi} \in \mathcal{E}_\pi$ whenever μ has full support in Z (see, for example, [6, Thm 3.2]). This fact together with [6, Cor 3.1] imply that $\hat{\pi}$ solves the reverse I-projection (7) and $\hat{\pi} = \pi_{\theta^*}$. ■

An important consequence of this lemma is that whenever Γ^n has full support over Z, the ML estimate $\hat{\theta}_n$ exists. This is because

$$\hat{\theta}_n := \underset{\theta\in\mathbb{R}^d}{\arg\max}\langle\Gamma^n, \pi_\theta\rangle = \underset{\theta\in\mathbb{R}^d}{\arg\min} D(\Gamma^n\|\pi_\theta). \tag{8}$$

The following theorem, the first part of which is known (see, for example, [6]), characterizes the asymptotic behavior of the statistic $D(\Gamma^n\|\pi_{\hat{\theta}^n})$. The main result of this paper is the second part which we prove in the appendix.

Theorem III.2. *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal $\mu \in \mathcal{P}(\mathsf{Z})$ with full support over Z. Let $\theta^* \in \mathbb{R}^d$ be as in Lemma III.1 and let $\pi_{\hat{\theta}_n}$ denote the ML estimate of the underlying distribution from the exponential family (2) based on the first n observations. Then we have,*

(i) *If $\mu \in \mathcal{E}_\pi$, then*

$$2nD(\Gamma^n\|\pi_{\hat{\theta}_n}) \xrightarrow[n\to\infty]{d.} \chi^2_{N-d-1} \tag{9}$$

where χ^2_a denotes a chi-squared random variable with a degrees of freedom.

(ii) *If $\mu \notin \mathcal{E}_\pi$, then*

$$\sqrt{n}(D(\Gamma^n\|\pi_{\hat{\theta}_n}) - D(\mu\|\pi_{\theta^*})) \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \sigma^2_\mu) \tag{10}$$

where $\mathcal{N}(0, \sigma^2)$ denotes a mean zero Gaussian random variable with variance σ^2 and $\sigma^2_\mu := \mathsf{Var}_\mu\left[\log\frac{\mu(Z)}{\pi_{\theta^*}(Z)}\right]$ denotes the variance of the random variable $\log\frac{\mu(Z)}{\pi_{\theta^*}(Z)}$ when $Z \sim \mu$. □

The above results suggest that for large enough n, we have the following approximations of the test statistic when $Z_i \sim \mu$

$$D(\Gamma^n\|\pi_{\hat{\theta}_n}) \approx \begin{cases} \frac{\chi^2_{N-d-1}}{2n} & \text{if } \mu \in \mathcal{E}_\pi \\ D(\mu\|\pi_{\theta^*}) + \frac{\mathcal{N}(0,\sigma^2_\mu)}{\sqrt{n}} & \text{if } \mu \notin \mathcal{E}_\pi. \end{cases} \tag{11}$$

The interesting property of the test statistic $D(\Gamma^n\|\pi_{\hat{\theta}_n})$ is the fact that the asymptotic distribution of this statistic is the same irrespective of which distribution π_θ is true under the null hypothesis $\mathcal{H}_0$. Thus the first approximation in (11) can be used to determine the threshold levels for the test (6) so as to meet an approximate false alarm probability constraint under $\mathcal{H}_0$. The second approximation of (11) on the other hand enables us to approximate the error performance of the test for alternate hypothesis distributions $\mu \notin \mathcal{E}_\pi$.

A. Goodness-of-fit with outliers

One of the problems in model-fitting problems is the presence of outliers in data. For example, in the problem described above, while most of the observations in the sequence $\boldsymbol{Z}$ may be drawn from some member π_θ of the exponential family, a small fraction of these points maybe outliers which do not correspond to the exponential family model. One approach to

model outliers is to assume that the true distribution of the observations is a mixture of the form

$$(1-\epsilon)\pi_\theta + \epsilon\xi \text{ where } \xi \in \mathcal{P}(\mathsf{Z}). \tag{12}$$

Here ξ is the unknown distribution of the outliers and $\epsilon \in [0,1]$ is the fraction of outliers in the data. Such distributions constitute an ϵ-contamination class [1]. In general the outlier distribution ξ is allowed to be arbitrary in $\mathcal{P}(\mathsf{Z})$ while in some cases a uniform distribution for the outliers can be justified.

We now obtain approximate expressions and bounds for the limiting divergence $D(\mu\|\pi_{\theta^*})$ and variance σ_μ^2 appearing in Theorem III.2 when μ is of the form (12) and ϵ is small. For ease of illustration we will use h to denote $\xi - \pi_\theta$ so that μ is now of the form $\pi_\theta + \epsilon h$, with $h \in \mathbb{R}^N$. We also define the $d \times N$ matrix Ψ as

$$\Psi(i,j) := \psi_i(z_j), \quad z_j \in \mathsf{Z}, \quad 1 \le i \le d$$

and use diag(v) to denote a square matrix with entries from vector v along its principal diagonal. The following proposition is proved in the appendix. In these results we use the standard big-O notation: for any function $g(\epsilon)$ the notation $\mathrm{O}(g(\epsilon))$ denotes some function $f(\epsilon)$ which satisfies the condition that there exists $\kappa > 0$ such that for ϵ small enough, we have $|f(\epsilon)| < \kappa g(\epsilon)$.

Proposition III.3. *Suppose $\mu \in \mathcal{P}(\mathsf{Z})$ is of the form (12) and let θ^* be as defined in Theorem III.2.*

(i) *The following approximations hold:*

$$D(\mu\|\pi_{\theta^*}) = \tfrac{1}{2}\epsilon^2 h^T G_\theta h + \mathrm{O}(\epsilon^3) \tag{13}$$

$$\sigma_\mu^2 = \epsilon^2 h^T G_\theta h + \mathrm{O}(\epsilon^3) \tag{14}$$

where

$$G_\theta = diag(\frac{1}{\pi_\theta}) - \Psi^T H_\theta^{-1}\Psi$$

with $H_\theta := \nabla^2\Lambda(\theta)$.

(ii) *The divergence and variance satisfy the following bounds*

$$D(\mu\|\pi_{\theta^*}) \le \log(1-\epsilon) + \epsilon + \frac{\epsilon^2\delta_\theta}{1-\epsilon} \tag{15}$$

$$\sigma_\mu^2 \le 2\epsilon^2\delta_\theta + \mathrm{O}(\epsilon^3). \tag{16}$$

where $\delta_\theta := (\min_z \pi_\theta(z))^{-1}$. □

Remark: A simpler version of the problem studied in this paper is when the family $\mathcal{E}_\pi$ in (5) is replaced by a single distribution π. Asymptotics of $D(\Gamma^n\|\pi)$ were studied in [5]. The results (9) and (10) continue to hold with $\pi_{\hat{\theta}_n}$ and π_θ^* replaced by π and with $d = 0$. In the presence of outliers, it can be shown that results (13) and (14) continue to hold with G_θ replaced by diag$(\frac{1}{\pi})$.

B. Implications of the results

In the rest of this section we list some of the applications of the results presented in this paper.

1) Power of rejecting alternate hypotheses: One of the important facets of model fitting that is often underemphasized is the ability of the goodness-of-fit procedure to reject wrong hypotheses. Suppose we are performing the test (6) and the true distribution μ of the observations $\{Z_i\}$ lies outside the exponential family $\mathcal{E}_\pi$. For such distributions we can use the second expression in (11) to approximate the probability of wrongly accepting the hypothesis $\mathcal{H}_0$ while performing the test (6).

2) Sensitivity to outliers: Another use of Theorem III.2 is to quantify the performance degradation of the goodness-of-fit test (6) when the observations are drawn from $\mathcal{E}_\pi$ but are corrupted by outliers. We see that when the true distribution $\mu = (1-\epsilon)\pi_\theta + \epsilon\xi \notin \mathcal{E}_\pi$, the test statistic no longer converges to zero, but instead converges to $D(\mu\|\pi_{\theta^*})$ with a standard deviation of order $\frac{1}{\sqrt{n}}\sigma_\mu$. Proposition III.3 illustrates how the divergence and variance vary as a function of ϵ. From the approximate expressions in (13) and (14) we can argue that those distributions $\pi_\theta \in \mathcal{E}_\pi$ with large eigenvalues for the corresponding matrix G_θ are most sensitive to outliers.

3) Robustifying for outliers: In practical scenarios when we expect to have outliers in our data, we may wish to make our goodness-of-fit test robust to outliers. In this case, we would like to expand our null hypothesis to

$$\mathcal{H}_0^\epsilon : Z_i \sim \text{i.i.d.}\,(1-\epsilon)\pi_\theta + \epsilon\xi, \text{ where } \theta \in \mathbb{R}^d, \xi \in \mathcal{P}(\mathsf{Z}).$$

If ϵ is small, then it may be reasonable to use the same statistic $D(\Gamma^n\|\pi_{\hat{\theta}_n})$ for goodness-of-fit as before. However, we may now reject hypothesis $\mathcal{H}_0^\epsilon$ only if the test statistic value cannot be explained by any distribution in $\mathcal{H}_0^\epsilon$. Proposition III.3(ii) gives us an exact bound on the divergence $D(\mu\|\pi_{\theta^*})$ and an approximate bound on the variance σ_μ^2 as a function of θ. If the parameters θ of interest belong to a compact subset Θ of $\mathbb{R}^d$ we know that $\max_{\theta\in\Theta}\delta_\theta$ is finite and hence these bounds can be used to choose the threshold τ in (6) to ensure that we approximately meet a false alarm constraint under all distributions of the form:

$$\mathcal{H}_0^{\epsilon,\Theta} : Z_i \sim \text{i.i.d.}\,(1-\epsilon)\pi_\theta + \epsilon\xi, \text{ where } \theta \in \Theta, \xi \in \mathcal{P}(\mathsf{Z}).$$

C. Choice of goodness-of-fit statistic

Consider the problem of testing the following simple null hypothesis

$$\mathcal{H}_0^\pi : Z_i \sim \text{i.i.d.}\,\pi, i = 1, 2, \ldots$$

where $\pi \in \mathcal{P}(\mathsf{Z})$. Hoeffding [7] proved that the test that uses the divergence statistic $D(\Gamma^n\|\pi)$ is universally optimal in an error exponent sense. This test maximizes the type-II error exponent (i.e. the error exponent under the alternate hypothesis $(\mathcal{H}_0^\pi)^c$) for all distributions subject to a constraint on the type-I error exponent (i.e. the error exponent under the null hypothesis $\mathcal{H}_0^\pi$). Now consider the problem of testing the following composite null hypothesis

$$\mathcal{H}_0^{\mathbb{P}} : Z_i \sim \text{i.i.d.}\,\mu, \quad \mu \in \mathbb{P} \tag{17}$$

where $\mathbb{P}$ is some subset of $\mathcal{P}(\mathsf{Z})$. This problem was studied in [2] and [4] when $\mathbb{P}$ is a linear family. It was shown in [2]

that a threshold test on $\inf_{\mu\in\mathbb{P}} D(\Gamma^n\|\mu)$ optimizes the type-II error exponent subject to a constraint on the worst-case type-I error-exponent. The composite hypothesis testing problem we study in (5) is identical to the problem in (17) with $\mathbb{P} = \mathcal{E}_\pi$. Furthermore, since the ML estimate $\hat{\theta}_n$ solves the reverse I-projection problem of (8) it follows that the test (6) is optimal in an error-exponent sense for solving (17) when $\mathbb{P} = \mathcal{E}_\pi$.

IV. Conclusion and future work

We have established the asymptotic behavior of the goodness-of-fit statistic with respect to an exponential family under general measures from the probability simplex. Our results can be used to approximate the power of the test to reject distributions from outside the exponential family. We have characterized the sensitivity of the goodness-of-fit test to data outliers and also provided guidelines for designing tests that are robust to outliers.

Although our results are for an exponential family of distributions, we believe that our approach can also be used to obtain similar results for general parametric classes of distributions on a finite alphabet. Another direction for future work is to analyze the implications of these results for goodness-of-fit testing using quantized observations drawn from parametric distributions on infinite alphabets. We are also seeking tighter bounds in Proposition III.3(ii) that do not explicitly depend on δ_θ.

Acknowledgements

This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006.

Appendix

A. Proof of Theorem III.2

Proof: Here we prove only the second part of the theorem since the first is well known (see, e.g., [6]). Since the reverse I-projection is attained, it follows that $D(\mu\|\pi_\theta)$ has a stationary point at $\theta^* = \theta^*(\mu)$. This means that we have,

$$\langle \mu, \psi\rangle = \nabla\Lambda(\theta^*). \tag{18}$$

This can be viewed as a fixed point equation in μ and $\theta(\mu)$. Since $\nabla^2\Lambda(\theta^*)$ is invertible, it follows by the Implicit Function Theorem [8] that $\theta^*(\mu)$ is continuously differentiable in a Euclidean neighborhood $N_\delta(\mu)$ of μ. It follows that whenever $\Gamma^n \in N_\delta(\mu)$, the ML estimate $\hat{\theta}_n$ is guaranteed to exist. Now since we know by the strong law of large numbers that $\Gamma^n \xrightarrow[n\to\infty]{a.s.} \mu$, we can argue via Slutsky's theorem [9] that it suffices to establish the weak convergence of

$$\sqrt{n}(D(\Gamma^n\|\pi_{\hat{\theta}_n}) - D(\mu\|\pi_{\theta^*}))\mathbb{I}\{\Gamma^n \in N_\delta(\mu)\}$$

where $\mathbb{I}\{E\}$ is the indicator function of event E.

Now consider the following decomposition:

$$D(\Gamma^n\|\pi_{\hat{\theta}_n}) - D(\mu\|\pi_{\theta^*}) = T_1 + T_2 \tag{19}$$

where $T_1 = D(\Gamma^n\|\pi_{\hat{\theta}_n}) - D(\Gamma^n\|\pi_{\theta^*})$ and $T_2 = D(\Gamma^n\|\pi_{\theta^*}) - D(\mu\|\pi_{\theta^*})$. For $\Gamma^n \in N_\delta(\mu)$, we have

$$T_1 = \langle \Gamma^n, (\hat{\theta}_n - \theta^*)^T\psi\rangle - \Lambda(\hat{\theta}_n) + \Lambda(\theta^*).$$

Using the ML condition $\langle \Gamma^n, \psi\rangle = \nabla\Lambda(\hat{\theta}_n)$, and a second order Taylor's expansion of $\Lambda(.)$ about $\hat{\theta}_n$, we have

$$\begin{aligned} T_1 &= (\hat{\theta}_n - \theta^*)^T\nabla\Lambda(\hat{\theta}_n) - \Lambda(\hat{\theta}_n) + \Lambda(\theta^*) \\ &= \tfrac{1}{2}(\hat{\theta}_n - \theta^*)^T\nabla^2\Lambda(\tilde{\theta}_n)(\hat{\theta}_n - \theta^*) \end{aligned} \tag{20}$$

where $\tilde{\theta}_n = \gamma\theta^* + (1-\gamma)\hat{\theta}_n$ with $\gamma = \gamma(n) \in [0,1]$. Now we know that $\hat{\theta}_n = \theta^*(\Gamma^n)$ holds when $\Gamma^n \in N_\delta(\mu)$. Thus by the differentiability of the θ^* function, we have

$$\hat{\theta}_n - \theta^*(\mu) = M_n(\Gamma^n - \mu)$$

where matrix M_n is given by $M_n(i,j) = \frac{\partial\theta_i^*(\tilde{\Gamma}_n)}{\partial\mu_j}$ where $\tilde{\Gamma}_n$ is some convex combination of Γ^n and μ. Thus (20) can be written as

$$T_1 = (\Gamma^n - \mu)^T M_n^T\nabla^2\Lambda(\tilde{\theta}_n)M_n(\Gamma^n - \mu).$$

Now as $n \to \infty$ we have by the continuity of $\nabla^2\Lambda(.)$ and differentiability of $\theta^*(.)$

$$M_n^T\nabla^2\Lambda(\tilde{\theta}_n)M_n\mathbb{I}\{\Gamma^n \in N_\delta(\mu)\} \xrightarrow[n\to\infty]{a.s.} Constant$$

Furthermore, we know by the central limit theorem that $n(\Gamma^n - \mu)(\Gamma^n - \mu)^T$ converges in distribution to a finite valued random matrix. Using these results and applying Slutsky's theorem we get

$$n^{\frac{1}{2}}T_1\mathbb{I}\{\Gamma^n \in N_\delta(\mu)\} \xrightarrow[n\to\infty]{d.} 0.$$

It follows from [5, Thm. III.3] that

$$\sqrt{n}(T_2 - D(\mu\|\pi_{\theta^*})) \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \sigma_\mu^2).$$

Combining the results on convergence of T_1 and T_2, we arrive at the desired result. ■

B. Proof of Proposition III.3

Proof of part (i): For any $\nu \in \mathcal{P}(\mathsf{Z})$ in the neighborhood of π_θ let $\theta^*(\nu)$ denote the reverse I-projection as before. We know that θ^* satisfies $\langle \nu, \psi\rangle = \nabla\Lambda(\theta^*(\nu))$. Furthermore, since $\nabla^2\Lambda(\theta^*)$ is invertible, we know by the Implicit Function Theorem that $\theta^*(\nu)$ is differentiable with respect to ν and if we define the $d \times N$ matrix M_θ by $M_\theta(i,j) := \frac{\partial\theta_i^*(\nu)}{\partial\nu_j}$ we have for all ν in some neighborhood of π_θ,

$$\Psi = \nabla^2\Lambda(\theta^*(\nu))M_{\theta^*(\nu)}. \tag{21}$$

For $\mu = \pi_\theta + \epsilon h$ we have the following approximation via Taylor's expansion of $\theta^*(\nu)$ about $\nu = \pi_\theta$:

$$\theta^*(\mu) = \theta + \epsilon M_\theta h + \mathrm{O}(\epsilon^2).$$

Now let θ^* denote $\theta^*(\mu)$. Extending the notation of measures as vectors to likelihood ratios we have,

$$\log\frac{\pi_{\theta^*}}{\pi_\theta} = \Psi^T(\theta^* - \theta) - (\Lambda(\theta^*) - \Lambda(\theta))\underline{1} \tag{22}$$

$$= (\Psi^T - \underline{1}(\nabla\Lambda(\theta))^T)(\theta^* - \theta) + \mathrm{O}(\epsilon^2)$$

$$= \epsilon(\Psi^T - \underline{1}(\nabla\Lambda(\theta))^T)M_\theta h + \mathrm{O}(\epsilon^2) \tag{23}$$

where $\underline{1}$ is an $N \times 1$ vector of all 1's. Now using the fact that $\nabla\Lambda(\theta^*) = \Psi\mu$ (see (18)), we obtain from (22):

$$\langle \mu, \log \frac{\pi_{\theta^*}}{\pi_\theta} \rangle = (\nabla\Lambda(\theta^*))^T(\theta^* - \theta) - (\Lambda(\theta^*) - \Lambda(\theta)). \quad (24)$$

We know by the second order Taylor's expansion of $\Lambda(.)$ that

$$\begin{aligned}\Lambda(\theta) &= \Lambda(\theta^*) + (\theta - \theta^*)^T \nabla\Lambda(\theta^*) \\ &\quad + \tfrac{1}{2}(\theta - \theta^*)^T \nabla^2\Lambda(\theta^*)(\theta - \theta^*) + \mathrm{O}(\epsilon^3).\end{aligned}$$

Using this relation in (24) and using H_θ for the Hessian, we get

$$\begin{aligned}&\langle \mu, \log \frac{\pi_{\theta^*}}{\pi_\theta} \rangle \\ &= \tfrac{1}{2}(\theta - \theta^*)^T H_{\theta^*}(\theta - \theta^*) + \mathrm{O}(\epsilon^3) \\ &= \tfrac{1}{2}\epsilon^2 h^T M_\theta^T H_{\theta^*} M_\theta h + \mathrm{O}(\epsilon^3) \\ &= \tfrac{1}{2}\epsilon^2 h^T \Psi^T H_\theta^{-1} H_{\theta^*} H_\theta^{-1} \Psi h + \mathrm{O}(\epsilon^3) \\ &= \tfrac{1}{2}\epsilon^2 h^T \Psi^T H_\theta^{-1} \Psi h + \mathrm{O}(\epsilon^3) \quad (25)\end{aligned}$$

where the last step follows by applying the analyticity of Λ to approximate H_{θ^*} with H_θ up to $\mathrm{O}(\epsilon)$. We know that

$$D(\mu\|\pi_{\theta^*}) = D(\mu\|\pi_\theta) - \langle \mu, \log \frac{\pi_{\theta^*}}{\pi_\theta} \rangle. \quad (26)$$

By Taylor's expansion of $D(\nu\|\pi_\theta)$ about $\nu = \pi_\theta$, we have,

$$\begin{aligned}D(\mu\|\pi_\theta) &= \tfrac{1}{2}(\mu - \pi_\theta)^T \mathrm{diag}(\frac{1}{\pi_\theta})(\mu - \pi_\theta) + \mathrm{O}(\epsilon^3) \\ &= \tfrac{1}{2}\epsilon^2 h^T \mathrm{diag}(\frac{1}{\pi_\theta}) h + \mathrm{O}(\epsilon^3).\end{aligned}$$

Combining this result with (25) using (26) we get (13).

For obtaining the approximate value of the variance $\sigma_\mu^2 = \mathsf{Var}_\mu[\log \frac{\mu}{\pi_{\theta^*}}(Y)]$ we first note that we have the following approximation for the log-likelihood ratio (LLR) function $L(.) := \log \frac{\mu}{\pi_{\theta^*}}(.)$:

$$\begin{aligned}L &= \log \frac{\pi_\theta + \epsilon h}{\pi_{\theta^*}} = \log \frac{\pi_\theta}{\pi_{\theta^*}} + \log(1 + \frac{\epsilon h}{\pi_{\theta^*}}) \\ &= -\epsilon(\Psi^T - \underline{1}(\nabla\Lambda(\theta))^T) M_\theta h + \frac{\epsilon h}{\pi_{\theta^*}} + \mathrm{O}(\epsilon^2)\end{aligned}$$

where the last step follows from (23) and the fact that $\log(1 + x) = x + \mathrm{O}(x^2)$. From the fact that the LLR function L is of order $\mathrm{O}(\epsilon)$, it can easily be seen that

$$\mathsf{Var}_\mu(L(Y)) = \mathsf{Var}_{\pi_\theta}(L(Y)) + \mathrm{O}(\epsilon^3). \quad (27)$$

The variance term on the right side can be expressed via the vector notation as

$$\mathsf{Var}_{\pi_\theta}(L(Y)) = \epsilon^2 B^T(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^T)B + \mathrm{O}(\epsilon^3) \quad (28)$$

where B is the vector L with the invariant part $\epsilon\underline{1}(\nabla\Lambda(\theta))^T M_\theta h$ omitted (since it does not contribute to the variance), and with $\frac{h}{\pi_{\theta^*}}$ replaced by $\frac{h}{\pi_\theta}$ (since it is correct up to $\mathrm{O}(\epsilon)$):

$$B := (\mathrm{diag}(\frac{1}{\pi_\theta}) - \Psi^T M_\theta)h.$$

Simplifying and using $\Psi(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^T)\Psi^T = \nabla^2\Lambda(\theta)$ by (4) we obtain,

$$\begin{aligned}B^T(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^T)B &= h^T(\mathrm{diag}(\frac{1}{\pi_\theta}) - \Psi^T M_\theta \\ &\quad - M_\theta^T\Psi + M_\theta^T\nabla^2\Lambda(\theta)M_\theta)h \\ &= h^T(\mathrm{diag}(\frac{1}{\pi_\theta}) - \Psi^T H_\theta^{-1}\Psi)h.\end{aligned}$$

where the last step follows from (21). Combining with (27) and (28) the result follows. ■

Proof of part (ii): We have

$$\begin{aligned}D(\mu\|\pi_{\theta^*}) &\le D(\mu\|\pi_\theta) = D((1-\epsilon)\pi_\theta + \epsilon\xi\|\pi_\theta) \\ &\le \max_z D((1-\epsilon)\pi_\theta + \epsilon 1_z\|\pi_\theta). \quad (29)\end{aligned}$$

where $1_z \in \mathcal{P}(\mathsf{Z})$ has unit mass on $z \in \mathsf{Z}$. The inequality (29) follows from the convexity of KL divergence. Now for any distribution $\nu \in \mathcal{P}(\mathsf{Z})$ we have,

$$\begin{aligned}&D((1-\epsilon)\nu + \epsilon 1_z\|\nu) \\ &= \sum_y ((1-\epsilon)\nu(y))\log(1-\epsilon) - (1-\epsilon)\nu(z)\log(1-\epsilon) \\ &\qquad + ((1-\epsilon)\nu(z) + \epsilon)\log(1 - \epsilon + \frac{\epsilon}{\nu(z)}) \\ &= (1-\epsilon)\log(1-\epsilon) + \epsilon\log(1-\epsilon) + \epsilon(x+1)\log(1 + \frac{1}{x}) \\ &= \log(1-\epsilon) + \epsilon(x+1)\log(1 + \frac{1}{x}) \quad (30)\end{aligned}$$

where $x = \frac{(1-\epsilon)\nu(z)}{\epsilon}$. Now using $\log(1 + \frac{1}{x}) \le \frac{1}{x}$, and substituting for x, we get

$$D((1-\epsilon)\nu + \epsilon 1_z\|\nu) \le \log(1-\epsilon) + \epsilon + \frac{\epsilon^2}{(1-\epsilon)\nu(z)}. \quad (31)$$

Combining with (29) we get (15). The second result (16) follows from the approximation (14) together with the fact that $h^T G_\theta h \le h^T \mathrm{diag}(\frac{1}{\pi_\theta})h \le h^T h \delta_\theta \le 2\delta_\theta$. ■

References

[1] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[2] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stoch. Proc. and Applns*, vol. 116, no. 5, pp. 724–756, May 2006.

[3] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. Hoboken, New Jersey: John Wiley & Sons Inc., 1987.

[4] J. Unnikrishnan, S. Meyn, and V. Veeravalli, "On thresholds for robust goodness-of-fit tests," in *Information Theory Workshop (ITW), 2010 IEEE*, September 2010, pp. 1–4.

[5] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1587 –1603, March 2011.

[6] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.

[7] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.

[8] D. Luenberger, *Linear and nonlinear programming*, 2nd ed. Norwell, MA: Kluwer Academic Publishers, 2003.

[9] P. Billingsley, *Convergence of Probability Measures*. New York: John Wiley & Sons, 1968.