Automated statistical tests for probabilistic programs

Feras Saad[†], Cameron Freer[†], Nate Ackerman[‡], Vikash Mansinghka[†]

[†]MIT, [‡]Harvard University

Problem Description: Goodness-of-Fit Testing

- Let $\mathbf{p}$ be a discrete distribution over discrete domain. Given data $y_{1:n} ::= \{y_1, \ldots, y_n\}$ drawn i.i.d. from unknown distribution $\mathbf{q}$, is there sufficient evidence to reject the hypothesis $\mathbf{p} = \mathbf{q}$?
- Many techniques apply to the case of continuous distributions.
- For low-dimensional distributions standard tests: Pearson chi-square, likelihood-ratio test, Kolmogorov-Smirnov.
- For high-dimensional distributions, standard statistics are:
  (i) intractable to compute;
  (ii) have little/no power;
  (iii) statistical assumptions typically not met (data too sparse).

observed data

$\mathbf{q} \sim^{\text{iid}} \{y_1, y_2, \ldots, y_n\}$

unknown distribution

$\mathbf{p}$

hypothesized distribution

is $\mathbf{p} = \mathbf{q}$?

Stochastic Rank Statistic

- New simulation-based statistic for goodness-of-fit testing high-dimensional data called the *stochastic rank statistic* (SRS).
- Main idea: For any elements y and y' in the domain, define a linear order $y \prec y'$. If $\mathbf{p} = \mathbf{q}$ then observed data point y_i $(i = 1, \ldots, n)$ should be uniformly distributed when ranked within a large dataset $\{y'_1, \ldots, y'_M\} \sim^{\text{iid}} \mathbf{p}$.
- Issue: For discrete data, order statistics are ill-defined (there are ties).
- Solution: Break ties uniformly-at-random by pairing each y_i with a uniform number u_i used to break ties.
- Exact: The SRS has an exact (non-asymptotic) null distribution. Easy to use for hypothesis testing (no approximations).
- Consistency: The SRS is uniformly distributed *if and only if* $\mathbf{p} = \mathbf{q}$.
- Practical: Goodness-of-fit test is simple and easy to implement.
- Flexible: Practitioner encodes domain-knowledge by designing linear order $\prec$ to specify which characteristics to check for goodness-of-fit.

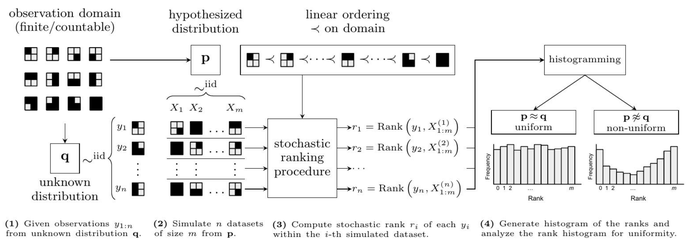
Discrete Goodness-of-Fit Tests Using Stochastic Rank Statistics

Algorithm 1 Discrete Goodness-of-Fit Testing Procedure

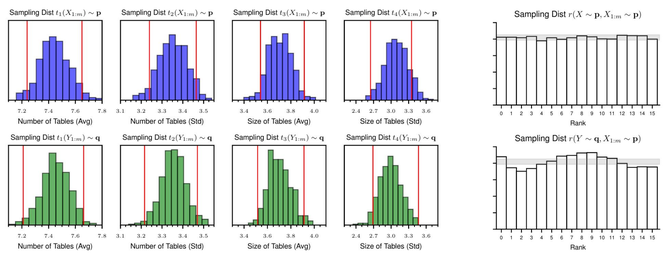
Input: simulator for candidate distribution $\mathbf{p}$ over finite or countable sample space T; observed samples $\{y_1, y_2, \ldots, y_n\}$ sampled i.i.d. from unknown distribution $\mathbf{q}$; strict total order $\prec$ on T, of any order type; number $m \geq 1$ of datasets to resimulate; significance level α;

Output: Decision whether to reject null hypothesis $\mathsf{H}_0 : \mathbf{p} = \mathbf{q}$ against alternative $\mathsf{H}_1 : \mathbf{p} \neq \mathbf{q}$ at significance level α.

1: **for** $i = 1, 2, \ldots, n$ **do**
2: $\quad X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)} \sim^{\text{iid}} \mathbf{p}$
3: $\quad U_0^{(i)}, U_1^{(i)}, \ldots, U_m^{(i)} \sim^{\text{iid}} \mathsf{Uniform}(0, 1)$
4: $\quad r_i \leftarrow \sum_{k=1}^{m} \mathbb{I}[X_k^{(i)} \prec y_i] + \mathbb{I}[X_k^{(i)} = y_i, U_k^{(i)} < U_0^{(i)}]$
5: Compute p-value of observed ranks $\{r_1, \ldots, r_n\}$ assuming cell labels $\{0, 1, 2, \ldots, m\}$ and cell probabilities $1/(m+1)$.
6: **return** reject if $p \leq \alpha$, else not reject.



Example: SRS vs. Bootstrapped Probe Statistics

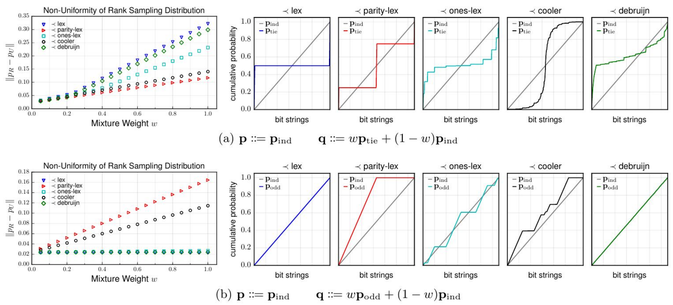


(a) Sampling distribution of four different probe statistics $\{t_1, t_2, t_3, t_4\}$ of a dataset of partitions, as sampled from $\mathbf{p}$ (Eq. (9); blue) and from $\mathbf{q}$ (Eq. (10); green) estimated by Monte Carlo simulation. Vertical red lines indicate 2.5% and 97.5% quantiles. Even though $\mathbf{p} \neq \mathbf{q}$, the distributions of these statistics are aligned in such a way that a statistic $t_j(Y_{1:m}) \sim \mathbf{q}$ is unlikely to appear as an extreme value in the sampling distribution of the corresponding statistic $t_j(X_{1:m}) \sim \mathbf{p}$, which leads to under-powered resampling-based tests.

(b) Monte Carlo simulation of the rank statistic illustrates its significant uniform distribution under the null hypothesis (top) and significant non-uniform distribution under the alternative hypothesis (bottom).

Figure 4: Comparison of the sampling distribution of (a) various bootstrapped probe statistics with (b) the SRS, for testing $\mathbf{p} ::= \mathsf{CRP}(0.26, 0.76)/2 + \mathsf{CRP}(0.19, 5.1)/2$ vs. $\mathbf{q} ::= \mathsf{CRP}(0.52, 0.52)$ (i.e., distributions on partitions of $\{1, \ldots, 20\}$).

Example: Distributions on Binary Strings



(a) $\mathbf{p} ::= \mathbf{p}_{\text{ind}} \quad \mathbf{q} ::= w\mathbf{p}_{\text{tie}} + (1-w)\mathbf{p}_{\text{ind}}$

(b) $\mathbf{p} ::= \mathbf{p}_{\text{ind}} \quad \mathbf{q} ::= w\mathbf{p}_{\text{odd}} + (1-w)\mathbf{p}_{\text{ind}}$

- Let domain $\{0,1\}^k$ be the set of all length k binary strings.
- Define the following distributions to be uniform over all strings $x = (x_1, \ldots, x_k) \in \{0,1\}^k$ which satisfy the given predicates:

$$\begin{aligned} \mathbf{p}_{\text{ind}} &: \text{ uniform on all strings,} \\ \mathbf{p}_{\text{tie}} &: x_1 = x_2 = \cdots = x_{k/2}. \\ \mathbf{p}_{\text{odd}} &: \textstyle\sum_{i=1}^{k} x_i \equiv 1 \pmod 2. \end{aligned}$$

- Each distribution assigns marginal probability $1/2$ to each bit x_i $(1 \leq i \leq k)$.
- All deviations from the uniform distribution $\mathbf{p}_{\text{ind}}$ are captured by higher-order relationships.
- The five orderings used for comparing binary strings are

$$\begin{aligned} \prec_{\text{lex}} &: \text{ Lexicographic (dictionary) ordering,} \\ \prec_{\text{par}} &: \text{ Parity of ones, ties broken using } \prec_{\text{lex}}, \\ \prec_{\text{one}} &: \text{ Number of ones, ties broken using } \prec_{\text{lex}}, \\ \prec_{\text{coo}} &: \text{ Cooler ordering (randomly generated),} \\ \prec_{\text{dbj}} &: \text{ De Bruijn sequence ordering.} \end{aligned}$$

- Null distribution: $\mathbf{p} ::= \mathbf{p}_{\text{ind}}$
- Alternative distributions: $\mathbf{q} ::= w\mathbf{p}_{\text{c}} + (1-w)\mathbf{p}_{\text{ind}}$ (mixtures of $\mathbf{p}_{\text{ind}}$ with the other two distributions)
- Bit strings of length $k = 16$ with $n = 256$ observations so that domain size is $65,536$ and 0.4% of the domain size is observed.

Theoretical Properties of the Stochastic Rank Statistic

Theorem 1. *Let $\mathcal{T}$ be a finite or countably infinite set, let $\prec$ be a strict total order on $\mathcal{T}$, let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions on $\mathcal{T}$, and let m be a positive integer. Consider the following random variables:*

$$X_0 \sim \mathbf{q} \tag{1}$$

$$X_1, X_2, \ldots, X_m \sim^{\text{iid}} \mathbf{p} \tag{2}$$

$$U_0, U_1, U_2, \ldots, U_m \sim^{\text{iid}} \mathsf{Uniform}(0,1) \tag{3}$$

$$R = \textstyle\sum_{j=1}^{m} \mathbb{I}[X_j \prec X_0] + \mathbb{I}[X_j = X_0, U_j < U_0]. \tag{4}$$

Then $\mathbf{p} = \mathbf{q}$ if and only if for all $m \geq 1$, the rank R is uniformly distributed on the set of integers $[m+1] ::= \{0, 1, 2, \ldots, m\}$.

Corollary 2. *If $\mathbf{p} \neq \mathbf{q}$, then there is some $m \geq 1$ such that R is not uniformly distributed on $[m+1]$.*

Theorem 3. *If $\mathbf{p} \neq \mathbf{q}$, then there is some $M \geq 1$ such that for all $m \geq M$, the rank R is not uniformly distributed on $[m+1]$.*

Corollary 4. *Let $\lhd$ denote the lexicographic order on $\mathcal{T} \times [0,1]$ induced by $(\mathcal{T}, \prec)$ and $([0,1], <)$. Suppose $\mathbb{E}\left[\mathbb{I}\left[(X, U_1) \lhd (Y, U_0)\right]\right] \neq 1/2$ for $Y \sim \mathbf{q}$, $X \sim \mathbf{p}$, and $U_0, U_1 \sim^{\text{iid}} \mathsf{Uniform}(0,1)$. Then for all $m \geq 1$, the rank R is non-uniformly distributed on $[m+1]$.*

Theorem 5. *Given significance level $\alpha = 2\Phi(-c)$ for $c > 0$, there is an ordering for which the proposed test with $m = 1$ achieves power $\beta \geq 1 - \Phi(-c)$ using*

$$n \approx 4c^2 / L_\infty(\mathbf{p}, \mathbf{q})^4 \tag{5}$$

samples from $\mathbf{q}$, where Φ is the distribution function of a standard normal.

Application: MCMC Convergence of Dirichlet Process Mixtures

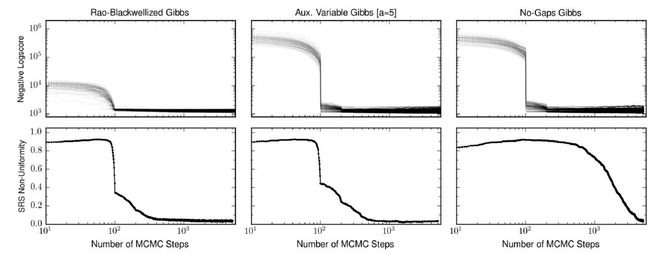


Figure : The uniformity of the SRS (bottom row) captures convergence behavior of MCMC sampling algorithms for Dirichlet process mixture models that are not captured by standard diagnostics such as the logscore (top row).

- In recent work, [TBS+18] describe general procedure for validating inference from Bayesian algorithms that can generate posterior samples.
- For prior $\pi(\theta)$ and likelihood $\pi(y|\theta)$, integrating the posterior over the joint returns the prior:

$$\pi(\theta) = \int \left[\pi(\theta|y')\pi(y'|\theta')\mathrm{d}y'\right] \pi(\theta')\mathrm{d}\theta'.$$

- Simulating datasets $y \sim \pi(y)$ from marginal distribution and running posterior inference $\pi(\theta|y)$ over parameters, the data-averaged posterior is identically distributed to prior.
- We repeatedly sampled $n = 100$ data points $\{x_1, \ldots, x_n\}$, simulated from DPMM over $\mathbb{R}^5$ with Gaussian component models.
- From SBC, the data-averaged posterior $\pi(z_{1:n}|x_{1:n})$ over $z_{1:n}$ is equivalent to the CRP prior $\pi(z_{1:n})$.
- Algorithm 2 specifies the ordering over partitions.
- Figure shows goodness-of-fit with respect to the true posterior of approximate samples $\hat{z}_{1:n}$ ($\approx 10^{115}$ different values) using Rao–Blackwellized Gibbs, Auxiliary Variable Gibbs, and No-Gaps Gibbs (Algorithms 3, 8, and 4 of [Nea00], respectively).

Algorithm 2 Total order $\prec$ on the set of partitions Π_N

Input: Partition $\pi ::= \{\pi_1, \pi_2, \ldots, \pi_k\} \in \Pi_N$ with k blocks. Partition $\nu ::= \{\nu_1, \nu_2, \ldots, \nu_l\} \in \Pi_N$ with l blocks.

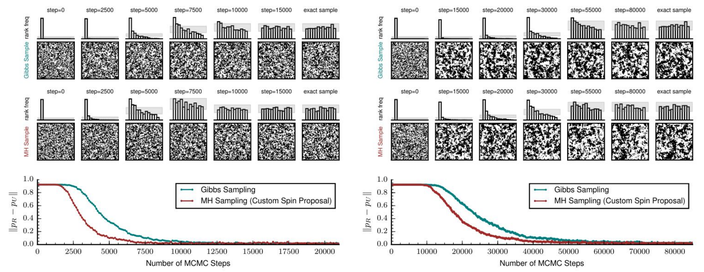
Output: LT if $\pi \prec \nu$; GT if $\pi \succ \nu$; EQ if $\pi = \nu$.

1: **if** $k < l$ **then return** LT ▷ ν has more blocks
2: **if** $k > l$ **then return** GT ▷ π has more blocks
3: $\bar{\pi} \leftarrow$ blocks of π sorted by value of least element in the block
4: $\bar{\nu} \leftarrow$ blocks of ν sorted by value of least element in the block
5: **for** $b = 1, 2, \ldots, l$ **do**
6: $\quad$ **if** $|\bar{\pi}_b| < |\bar{\nu}_b|$ **then return** LT ▷ $\bar{\nu}_b$ has more elements
7: $\quad$ **if** $|\bar{\pi}_b| > |\bar{\nu}_b|$ **then return** GT ▷ $\bar{\pi}_b$ has more elements
8: $\quad \pi'_b \leftarrow$ values in $\bar{\pi}_b$ sorted in ascending order
9: $\quad \nu'_b \leftarrow$ values in $\bar{\nu}_b$ sorted in ascending order
10: $\quad$ **for** $i = 1, 2, \ldots, |\pi'_b|$ **do**
11: $\quad\quad$ **if** $\pi'_{b,i} < \nu'_{b,i}$ **then return** LT ▷ π'_b has smallest element
12: $\quad\quad$ **if** $\pi'_{b,i} > \nu'_{b,i}$ **then return** GT ▷ π'_b has smallest element
13: **return** EQ

[TBS+18] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint*, (arXiv:1804.06788), 2018.

[Nea00] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Application: MCMC Convergence of Ising Model Samplers



(a) Temperature $T = 8$

(b) Temperature $T = 3$

Figure : Assessing the goodness-of-fit of approximate samples of a 64×64 Ising model for Gibbs sampling and Metropolis–Hastings sampling (with the custom spin proposal from [Mac03]) at two temperatures using the SRS. In both cases, the SRS converges to its uniform distribution more rapidly for samples obtained from MH than for those from Gibbs sampling.

This work was published as:

A Family of Exact Goodness-of-Fit Tests for High-Dimensional Discrete Distributions,
Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR 89:1640-1649, 2019.