Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling

Rina Foygel Barber* and Lucas Janson†

Abstract

Goodness-of-fit (GoF) testing is ubiquitous in statistics, with direct ties to model selection, confidence interval construction, conditional independence testing, and multiple testing, just to name a few applications. While testing the GoF of a simple (point) null hypothesis provides an analyst great flexibility in the choice of test statistic while still ensuring validity, most GoF tests for composite null hypotheses are far more constrained, as the test statistic must have a tractable distribution over the entire null model space. A notable exception is *co-sufficient sampling* (CSS): resampling the data conditional on a sufficient statistic for the null model guarantees valid GoF testing using any test statistic the analyst chooses. But CSS testing requires the null model to have a compact (in an information-theoretic sense) sufficient statistic, which only holds for a very limited class of models; even for a null model as simple as logistic regression, CSS testing is powerless. In this paper, we leverage the concept of *approximate sufficiency* to generalize CSS testing to essentially any parametric model with an asymptotically-efficient estimator; we call our extension "approximate CSS" (aCSS) testing. We quantify the finite-sample Type I error inflation of aCSS testing and show that it is vanishing under standard maximum likelihood asymptotics, for any choice of test statistic. We apply our proposed procedure both theoretically and in simulation to a number of models of interest to demonstrate its finite-sample Type I error and power.

Keywords: Goodness-of-fit test, approximate sufficiency, co-sufficiency, conditional randomization test, model-X, conditional independence testing, high-dimensional inference.

1 Introduction

Suppose we observe data X belonging to some sample space $\mathcal{X}$, and would like to test whether it comes from some parametric null model $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$, versus a more complex (usually higher-dimensional) model. This problem of so-called "goodness-of-fit" (GoF) testing is one of the most fundamental in statistics, with a vast literature exhibiting applications and theoretical and methodological development. We pause here to highlight a few of the many areas of statistics to which GoF testing is directly applicable, including some problems that are not obviously or commonly associated with GoF.

*Department of Statistics, University of Chicago
†Department of Statistics, Harvard University

Problem Domain 1 (Standard goodness-of-fit testing)**.** *GoF testing is commonly used to test a postulated model or distributional property, often as a precursor to further statistical analysis that assumes the postulated model/property to be correct. Such null models/properties include standard distributional families, nonparametric properties such as symmetry or log-concavity, time-series properties such as stationarity, and relational properties such as independence.*

Problem Domain 2 (Model selection)**.** *GoF testing can also be used to select a best-fitting model through simultaneously testing a family of models. For instance, this could be choosing a sparse model in regression, selecting the number of clusters or principle components in unsupervised learning, or identifying change points in a time series.*

Problem Domain 3 (Confidence interval construction)**.** *Suppose the data X is distributed according to a known model $\{P_{\gamma,\theta} : (\gamma, \theta) \in \Gamma \times \Theta\}$, where $\Gamma \subseteq \mathbb{R}^m$ and $\Theta \subseteq \mathbb{R}^d$, and the goal is to construct a confidence region for γ in the presence of the nuisance parameter θ. If for any γ_0, we can construct a GoF test for the null model $\{P_{\gamma_0,\theta} : \theta \in \Theta\}$, then the set of γ_0 at which we fail to reject the test constitutes a valid confidence region.*

Problem Domain 4 (Conditional independence testing)**.** *In many regression and graphical modeling problems, the primary question of interest is whether a given triple of random variables (X, Y, Z) satisfies conditional independence, i.e., $Y \perp\!\!\!\perp X \mid Z$. If $X \mid Z$ is distributed according to a known model $\{P_\theta(\cdot \mid Z) : \theta \in \Theta\}$, then testing conditional independence can be formulated as a GoF test where $\{P_\theta(\cdot \mid Z) : \theta \in \Theta\}$ is the null model for $X \mid Z, Y$. (Note that, when we apply a GoF test to the conditional independence problem in this way, we implicitly treat Y and Z as fixed, and check for goodness-of-fit of X's conditional model.)*

Problem Domain 5 (Multiple testing)**.** *In multiple testing, the goal is to reject a subset of a fixed family of null hypotheses. When each hypothesis test is a GoF test (i.e., its null is lower-dimensional than its alternative), testing any intersection of null hypotheses (i.e., testing the "global null" for any subset of hypotheses) constitutes a GoF test as well, and combining such intersection GoF tests through a closed testing procedure [Marcus et al., 1976] produces a subset to reject that controls the familywise error rate.*

The key challenges of any GoF testing problem are to find a test that is valid, in that it controls the Type I error at a prespecified significance level, and that is powerful, in that it rejects the null model as often as possible when it does not hold. For parametric null models (the focus of this paper), there are many existing methods for testing GoF, with canonical choices including the popular score, likelihood ratio, and Wald tests. The standard approach for these tests and many others is to prescribe a test statistic (chosen to be powerful under a given alternative model) and derive a (often asymptotic) null distribution for it. Such tests require certain regularity conditions on the alternative model (in order to construct a well-behaved test statistic) and on the null model (in order to establish the validity of the null distribution for the test statistic) that are generally quite similar to those needed for the maximum likelihood estimator under both the null and alternative to be asymptotically normal. While these tests are extremely popular and have been fruitfully applied through much of the history of statistics to many problems, the regularity assumptions placed on the alternative distribution in particular limit the ability to fully leverage domain knowledge to

maximize the statistical power. To elaborate, consider the following cases which often arise in practice when applying parametric GoF tests.

- **Some prior information is available about the relative plausibility of different regions of the alternative model.** Ideally we would like to incorporate this prior information into our test statistic in order to maximize power (e.g., through a test statistic derived from Bayesian inference), but standard GoF tests only provide the null distribution for a test statistic which is determined by the entire alternative space, and give little flexibility to incorporate prior knowledge into that test statistic while still retaining a valid null distribution. An extreme case would be when certain regions of the alternative are known to be completely implausible, i.e., we would like to remove them from the alternative model entirely, yet removing them from the model would violate the regularity conditions required for the alternative model. For example, we may know that under the alternative some k-dimensional parameter has at most $d < k$ non-zero entries, but we do not know which ones. Such a sparse alternative model would violate the usual assumption that the parameter space is convex, forcing one to ignore the sparsity and instead operate under (and hence derive a test statistic from) the larger, mostly implausible, k-dimensional alternative model. As we see in the next scenario, if k is too large, even this route is not feasible.

- **The alternative model is high- or even infinite-dimensional.** Since standard GoF tests treat their prescribed test statistics as fixed in the asymptotic regime in which they prove validity, those test statistics can only be designed to be powerful against fixed- (and finite-) dimensional alternatives. If we have a high-dimensional alternative (i.e., whose dimension is not assumed negligible relative to the sample size, which includes any nonparametric alternative), we would ideally like to choose a test statistic which changes with the sample size to be powerful against a sequence of alternatives which changes as the sample size grows asymptotically. But standard GoF tests cannot accommodate such a choice, forcing one to instead operate under a fixed-dimensional alternative that may represent a vanishing fraction of plausible alternatives, or a very coarse approximation to the space of realistic alternatives.

The common thread in these cases is that the test statistic that would be most powerful to use given the domain knowledge at hand is not accompanied by theoretical guarantees or any known (exact or approximate) null distribution. In our simulations in Section 4, we will study some examples where standard tests can be applied (and will compare aCSS to the score test for those examples), and others where, as in the scenarios above, standard tests cannot be applied and thus a more flexible method like aCSS is necessary.

However, constructing a valid test around an arbitrary test statistic $T(X)$ is possible only in very limited settings. In particular, if Θ is simple, i.e., it contains only a single point so that there are no unknown/nuisance parameters in the null model, then any test statistic $T(X)$'s null distribution can be arbitrarily-well approximated computationally by repeatedly independently sampling copies $\widetilde{X}$ of X from the null distribution and recomputing the same test statistic on the copies. To be concrete, if the statistic $T(X)$ is chosen such that larger (positive) values are seen as evidence against the null, we can draw M i.i.d. copies

$\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ from the null distribution, and define a (discretized) p-value

$$\text{pval} = \text{pval}_T(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(\widetilde{X}^{(m)}) \geq T(X)\right\}\right), \tag{1.1}$$

which is guaranteed to satisfy $\mathbb{P}(\text{pval} \leq \alpha) \leq \alpha$ under the null, for any predefined rejection level α.

More generally, when Θ is not a singleton set (i.e., the null hypothesis is composite), in principle we can still construct a p-value of the form (1.1) as long as we are able to sample a set of copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ of X so that $X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are *exchangeable* under the null. We emphasize that this exchangeability property continues to enable the analyst to use any desired test statistic $T(X)$, as the validity of the p-value is unaffected. Of course, to achieve high power, we should aim to choose a function $T(X)$ that is likely to be large under the alternative hypothesis. Note that we absorb everything that is not X into the definition of the function T, e.g., for testing conditional independence $X \perp\!\!\!\perp Y \mid Z$ as in Problem Domain 4, T can depend arbitrarily on Y and Z as well (since, after conditioning on Y and Z, they are treated as fixed and nonrandom).

To summarize, we have seen that

> The problem of GoF testing with arbitrary test statistics can be reduced to one of sampling copies of X that are exchangeable under the null.

These copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ then act as a "control group" for the real data X, and we can compare the real statistic $T(X)$ against the "control group" values $T(\widetilde{X}^{(m)})$ to test the null. Of course, aside from the setting of a simple null, sampling exchangeable copies is not necessarily a straightforward task. In particular, in order to have power, we must ensure our null-exchangeable copies do not remain exchangeable under the alternative; for instance sampling $\widetilde{X}^{(1)} = \cdots = \widetilde{X}^{(M)} = X$ trivially satisfies the exchangeability property under the null, but also under any alternative, and hence clearly equation (1.1) produces a useless p-value of 1 with probability 1 (for any choice of T).

One way to sample exchangeable copies when Θ is composite is by conditioning on a sufficient statistic $S(X)$ for θ, since then by definition the conditional distribution $X \mid S(X)$ does not depend on θ. By drawing the copies $\widetilde{X}^{(m)}$ from this conditional distribution, we achieve exchangeability of $X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ (more concretely, X and its copies are i.i.d. conditional on $S(X)$). This approach is known as *co-sufficient sampling* (CSS) [Stephens, 2012]. However, this approach is viable in only a limited range of settings. In particular, many null models do not admit a compact (in an information-theoretic sense) sufficient statistic, meaning any sufficient statistic for the null model will remain sufficient for many alternative models as well. In such cases, which we term *degenerate*, CSS testing runs into the problem described at the end of the preceding paragraph—the copies $\widetilde{X}^{(m)}$ will still be exchangeable with X under the alternative, resulting in a powerless test. This situation arises quite often, and we will give a number of common examples shortly in Section 1.2.

As an alternative approach, we might consider the *parametric bootstrap* [Efron and Tibshirani, 1994], where after estimating the true parameter θ via some estimator $\widehat{\theta}$ (e.g., the maximum likelihood estimate), the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are sampled from $P_{\widehat{\theta}}$. While this

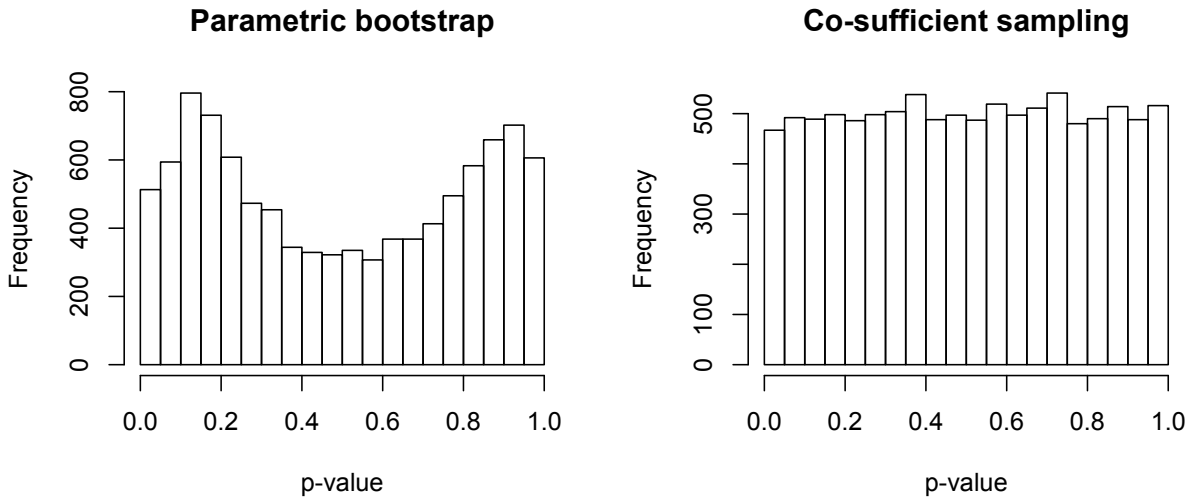

Figure 1: Comparison of the parametric bootstrap versus CSS, in a Gaussian linear model example where the null hypothesis is true. We can see that CSS yields uniformly distributed p-values, but the parametric bootstrap does not, resulting in an inflated Type I error rate. (See Section 1 for details.)

widely-used approach often works well in practice, the parametric bootstrap does not create exchangeable copies of the data, and is not guaranteed to achieve the desired Type I error level when paired with an arbitrary test statistic T—in fact, it may even lead to substantial error inflation. To take a simple example, consider a Gaussian linear regression setting where P_θ is given by the distribution $X \sim \mathcal{N}(\theta \cdot Z, \mathbf{I}_d)$, where $Z \in \mathbb{R}^n$ is a fixed covariate. Suppose we are interested in testing whether X in fact has more dependence with another covariate $Y \in \mathbb{R}^n$, and so our test statistic is given by

$$T(X) = \frac{(X^\top Y)^2}{(X^\top Z)^2}.$$

Figure 1 compares the parametric bootstrap against co-sufficient sampling (full details for this simulation are given in Appendix E). The results show that CSS results in a uniform distribution of p-values, while the parametric bootstrap results in a highly non-uniform distribution, and could lead to substantially inflated Type I error. Therefore, we would instead prefer to extend the CSS framework in order to enjoy theoretically guaranteed error control.

1.1 Our contribution

In this paper, we demonstrate how to escape the problem of zero power in the degenerate setting, by introducing a new generalization of CSS testing that conditions only on an approximately sufficient statistic [Le Cam, 1960, Van der Vaart, 2000, Le Cam, 2012]. We

call such a test an "approximate co-sufficient sampling" test. This paper makes four main contributions:

1. We propose *approximate co-sufficient sampling* (aCSS), which samples approximately exchangeable copies of the data by conditioning on an approximately sufficient statistic and plugging in a consistent estimator for the unknown parameter.

2. Under weak conditions closely resembling those for standard maximum likelihood asymptotics, we provide a finite-sample upper-bound for the total variation (TV) distance from exchangeability of our aCSS samples.

3. We show that the aforementioned TV bound translates directly to a bound on the Type I error inflation of an aCSS test that holds uniformly over the choice of test statistic, and we apply this bound to a number of important models to prove the inflation vanishes asymptotically as the sample size approaches infinity.

4. We provide general algorithms for aCSS and demonstrate their use in a series of simulations that exhibit the validity and power of aCSS testing.

1.2 Applications

The problem of zero power for CSS testing arises surprisingly often—while we call such settings "degenerate", they are not extreme cases but rather constitute a large portion of common statistical models of interest. To illustrate this, we will consider the following settings in which CSS testing is powerless, while aCSS testing can still be quite powerful and remains asymptotically valid for any test statistic. (Our theoretical results later on will quantify its finite-sample Type I error).

Model Class 1 (Data with associated covariates)**.** *Suppose $X = (X_1, \ldots, X_n)$ where the X_i's are independent, and each X_i has an associated covariate Z_i (i.e., the null distribution of each X_i depends on this Z_i). In this setting, the distribution of X (i.e., the joint distribution of $X_1, \ldots, X_n$) will often have the data X itself as a minimal sufficient statistic. This is even true when $X_i \,|\, Z_i$ follows logistic regression: for generic values of the Z_i's (e.g., if each $Z_i \in \mathbb{R}^d$ is drawn from some continuous distribution), the minimal sufficient statistic is equivalent to X itself because the value of $Z^\top X \in \mathbb{R}^d$ determines $X \in \{0,1\}^n$ uniquely, and hence is sufficient under any alternative as well (and hence any CSS test must be powerless). This problem class applies not just for GoF testing for conditional models for X (including the logistic model), but also for conditional independence testing as described in Problem Domain 4.*

Model Class 2 (Curved exponential families)**.** *Consider a null model that is a curved exponential family, i.e., a full-rank k-parameter exponential family with an added constraint that reduces the dimension of the parameter space to some $d < k$ and is nonlinear in the canonical parameters. In this setting, the minimal sufficient statistic is generally the same as that for the unconstrained (full-rank) exponential family. This means that any CSS test must be powerless against any alternative that lies in the larger exponential family, for example, if we want to test whether the parameter constraint holds or not. A classical example is*

the Behrens–Fisher problem of testing equality of (unknown) means between independent normal samples having different (unknown) variances: any CSS test will be powerless for every alternative pair of means and variances. The same issue arises in, e.g., the study of contingency tables (where the canonical family is multinomial and the null hypothesis imposes a constraint on its probabilities), and spatial and time-series models (where the null hypothesis imposes a spatial or temporal structure on the canonical parameters of an exponential family).

Model Class 3 (Heavy-tailed models)**.** *Many heavy-tailed models are not exponential families and do not admit compact sufficient statistics. For instance, the Cauchy location family's minimal sufficient statistic is given by the order statistics. Any CSS test of this null is therefore powerless against any i.i.d. alternative, since the order statistics are sufficient for this alternative as well. As another example, the Student's t scale family's minimal sufficient statistic is the order statistics of the absolute values, so any CSS test for it must be powerless against any i.i.d. symmetrical alternative.*

Model Class 4 (Models with latent variables)**.** *Many popular models capture domain-specific properties through latent variables. In such models, if we condition on the latent variable, then the data often comes from a well-behaved distribution with compact minimal sufficient statistic. However, when the latent variable is unobserved, we are forced to perform inference unconditionally, and the unconditional model rarely has a compact minimal sufficient statistic. Examples include hidden Markov models, mixture distributions, data with missing values, errors-in-variables models, and factor models.*

Later on in Section 4, we will return to Model Classes 1, 2, and 3 and give concrete examples of models where aCSS can be applied. We leave Model Class 4 for future work.

1.3 Related work

GoF testing dates back to the very early days of the field of statistics, and the literature even on just parametric GoF is far too numerous to cite. Instead, we focus our literature review on the subfield of CSS testing, which distinguishes itself within the broader field of parametric GoF testing by guaranteeing Type I error control with *any* test statistic under a parametric null model, the potential advantages of which have been described earlier in this section. In fact, some of the most foundational nonparametric tests can be thought of as CSS tests, including the permutation test (conditions on the order statistics for an i.i.d. null). The formal idea of a CSS test seems to date back to at least Bartlett [1937], although the value of sufficient statistics for GoF testing in the presence of nuisance parameters has also been used in many other ways, e.g., Durbin [1961], Kumar and Pathak [1977], Bell [1984] decompose the data into a minimal sufficient statistic and an ancillary statistic and construct a GoF test based on the parameter-free distribution of the ancillary statistic.

CSS testing has gained substantial interest in the last 30 years, though with a focus on non-degenerate hypothesis testing settings [Agresti, 1992, Engen and Lillegård, 1997, Agresti, 2001, O'Reilly and Gracia-Medrano, 2006, Lockhart et al., 2007, 2009, Lindqvist and Rannestad, 2011, Broniatowski and Caron, 2012, Lockhart, 2012, Stephens, 2012, Lindqvist and Taraldsen, 2013, Hazra, 2013, Beltrán-Beltrán and O'Reilly, 2019, Santos and Filho,

2019, Contreras-Cristán et al., 2019]. Our work differs from the existing work in CSS testing by allowing for degenerate (and non-degenerate) models by conditioning only on an approximately sufficient statistic. Similar techniques have been used to obtain exact confidence intervals in the presence of nuisance parameters [Lillegård and Engen, 1999], again in non-degenerate settings. As an example, Rosenbaum [1984], Kolassa [2003] study conditional independence testing of $X \perp\!\!\!\perp Y \mid Z$ where $X \mid Z$ follows a logistic regression model and Z is discrete; logistic regression is degenerate when Z has a continuous distribution but non-degenerate when Z is discrete.

For conditional independence testing (Problem Class 4), in the setting where H_0 is a simple null hypothesis (i.e., $X \mid Z$ has a known distribution), Candès et al. [2018] study procedures of the form (1.1) under the name "conditional randomization test"; their work also constructs the model-X knockoffs framework for simultaneously testing multiple conditional independence hypotheses (i.e., variable selection in a multivariate regression). This construction provides an exact "swap-exchangeability" property that enables variable selection, i.e., simultaneous conditional independence testing of many covariates when the multivariate covariates X come from a non-degenerate model, and leads to exact false discovery rate control [Barber and Candès, 2015]. Generalizing to the setting where H_0 is not simple, Huang and Janson [2020+] construct model-X knockoffs [Candès et al., 2018] conditional on a sufficient statistic, retaining the exact "swap-exchangeability" property and exact false discovery rate control; we can think of this as a knockoffs-analogue of CSS testing.

Moving beyond CSS testing, we are only aware of a few works which take a similarly approximate approach to that of the present paper. First, the most related to our approach is the work of Lillegård [2001], where they mention the possibility of an aCSS-type test to solve the Behrens–Fisher problem (i.e., testing for a difference of means between two Gaussian samples, described above in Model Class 2), but conclude that such an approach would be computationally intractable; they instead propose a heuristic sampling procedure which they support with simulations but no theory. Second, Kalbfleisch and Sprott [1970], Cox and Reid [1987] focus on parametric likelihood-based testing in the presence of nuisance parameters, but study the case where the nuisance parameters are orthogonal or asymptotically orthogonal to the parameter of interest. Finally, approximate Bayesian computation, also known as likelihood-free inference, conducts Bayesian inference conditional on a compact non-sufficient statistic, but for computational, as opposed to statistical reasons, since in the Bayesian framework there is no statistical downside to conditioning on as much as possible (see, e.g., Kousathanas et al. [2016] for such a paper that explicitly addresses the role of sufficiency).

1.4 Notation

We will write $\|\cdot\|$ to denote the usual Euclidean (ℓ_2) norm on vectors, and to denote the operator norm (i.e., spectral norm) on matrices. For a matrix M, $\lambda_{\max}(M)$ denotes its largest eigenvalue in the positive direction. We write $(x)_+$ to denote $\max\{x, 0\}$. We will write $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ to denote expectation or probability taken with respect to $X \sim P_\theta$, where the parametric family $\{P_\theta : \theta \in \Theta\}$ is our null model.

2 Method

The goal of approximate co-sufficient sampling is to generate copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ of the observed data X such that if the null hypothesis

$$H_0 : X \sim P_\theta \text{ for some } \theta \in \Theta \tag{2.1}$$

is true, then the random variables $X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are approximately exchangeable. Recalling the p-value defined in (1.1), we can then test the null hypothesis using any desired test statistic $T(X)$. The choice of statistic is completely unconstrained, and this flexibility enables us to design very powerful tests in many settings. Note that, although this flexibility allows us to design any form of function T, T itself (as a function, i.e., before seeing its argument) cannot depend on X. For example, if $T(X)$ uses X to tune parameters in a neural network and then computes a statistic of that neural network applied to X, then $T(\widetilde{X}^{(m)})$ cannot compute a statistic on the same (X-tuned) neural network applied to $\widetilde{X}^{(m)}$, but must use $\widetilde{X}^{(m)}$ to tune the parameters of a new neural network and compute a statistic of that ($\widetilde{X}^{(m)}$-tuned) neural network applied to $\widetilde{X}^{(m)}$.

To quantify our goal of generating approximately exchangeable copies of the data, we begin by defining a "distance to exchangeability":

Definition 1. *For any integer $k \geq 1$ and any set of random variables $A_1, \ldots, A_k$ with a joint distribution, define*

$$\mathsf{d}_{\mathsf{exch}}(A_1, \ldots, A_k) = \inf \left\{ \mathsf{d}_{\mathsf{TV}}\big((A_1, \ldots, A_k), (B_1, \ldots, B_k)\big) : B_1, \ldots, B_k \text{ are exchangeable} \right\}.$$

Here d_{TV} denotes the total variation distance, and the infimum is taken over all sets of k random variables $B_1, \ldots, B_k$ with an exchangeable joint distribution.

Of course, if $A_1, \ldots, A_k$ are exchangeable, then $\mathsf{d}_{\mathsf{exch}}(A_1, \ldots, A_k) = 0$. When we say informally that variables $A_1, \ldots, A_k$ are "approximately exchangeable", we mean that the distance to exchangeability is small.

Now we will see how this distance $\mathsf{d}_{\mathsf{exch}}$ relates to the problem of testing the null hypothesis (2.1) (Berrett et al. [2019] use a similar argument in a permutation test setting). Fix a threshold $\alpha \in [0, 1]$ and a function $T : \mathcal{X} \to \mathbb{R}$ (the test statistic). For any exchangeable random variables $(B_0, \ldots, B_M)$, by definition of exchangeability we have $\mathbb{P}\left(\text{pval}_T(B_0, B_1, \ldots, B_M) \leq \alpha\right) \leq \alpha$, and therefore,

$$\mathbb{P}\left(\text{pval}_T(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq \alpha\right) \leq \alpha + \mathsf{d}_{\mathsf{TV}}\big((X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}), (B_0, B_1 \ldots, B_M)\big).$$

Taking an infimum over all exchangeable distributions on $(B_0, B_1 \ldots, B_M)$, we have shown that

$$\mathbb{P}\left(\text{pval}_T(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq \alpha\right) \leq \alpha + \mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$$

under the null hypothesis H_0.

Therefore, we can see that, if we are able to construct copies of the data such that $\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ is small, then we can construct an approximately-valid test of H_0 using any desired test statistic T. From this point on, then, our task is to determine how we can use approximate sufficiency to generate such copies.

2.1 Overview

Consider any function $S = S(X)$ of the data, which is sufficient under the null hypothesis that $X \sim P_\theta$ for some $\theta \in \Theta$. Let $P(\cdot \,|\, s)$ be the conditional distribution of X given $S = s$ (sufficiency of $S(X)$ ensures that this distribution does not depend on θ). As described in Section 1, the co-sufficient sampling (CSS) method operates by drawing copies from this conditional distribution. That is, the joint distribution of the data and the copies, under the CSS method, is given by:

$$\begin{cases} X \sim P_{\theta_0}, \\ S = S(X), \\ \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \mid X, S \overset{\text{iid}}{\sim} P(\cdot \,|\, S), \end{cases}$$

where θ_0 is the unknown true parameter. Clearly, the real and fake data $X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are i.i.d. conditional on S, and are therefore exchangeable, meaning that the $\widetilde{X}^{(m)}$'s provide a valid "control group" for the real data X regardless of the unknown θ_0.

As discussed above, this framework is limited to only certain specific models, since many common models are "degenerate" (such as the model classes described in Section 1.2), where any sufficient statistic $S = S(X)$ reveals so much information about X that it leads to a completely powerless procedure against the alternative hypothesis of interest. We can instead consider statistics $S = S(X)$ that are not sufficient, but are *approximately sufficient*, meaning that the distribution $P_\theta(\cdot \,|\, S)$ is *approximately* unaffected by the value of θ—more concretely, if we can estimate θ with a consistent estimator $\widehat{\theta}$, then we only need to ensure that $P_{\widehat{\theta}}(\cdot \,|\, S) \approx P_\theta(\cdot \,|\, S)$. In fact, for many settings, maximum likelihood estimation is known to provide an asymptotically sufficient statistic [Le Cam, 1960, Van der Vaart, 2000, Le Cam, 2012]. Thus, we can take $S = S(X)$ to simply be $\widehat{\theta}_{\text{MLE}}(X)$, or more generally, any other estimator of θ_0 that is asymptotically sufficient.

In this setting, we write $P_{\theta_0}(\cdot \,|\, \widehat{\theta})$ to denote the conditional distribution of $X \,|\, \widehat{\theta}$ when the data is distributed as $X \sim P_{\theta_0}$ and we calculate $\widehat{\theta} = \widehat{\theta}_{\text{MLE}}(X)$. Of course, we cannot draw the copies from this distribution since θ_0 is unknown, but if $\widehat{\theta} = \widehat{\theta}_{\text{MLE}}(X)$ is approximately sufficient, then the distribution $P_{\theta_0}(\cdot \,|\, \widehat{\theta})$ should depend only slightly on θ_0. In particular, we will use $\widehat{\theta}$ itself as a plug-in estimate for θ_0, leading to the joint model

$$\begin{cases} X \sim P_{\theta_0}, \\ \widehat{\theta} = \widehat{\theta}_{\text{MLE}}(X), \\ \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \mid X, \widehat{\theta} \overset{\text{iid}}{\sim} P_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}). \end{cases}$$

These copies form an approximately-valid control group as long as $P_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}) \approx P_{\theta_0}(\cdot \,|\, \widehat{\theta})$.

In our aCSS algorithm, we will replace the deterministic step $\widehat{\theta} = \widehat{\theta}(X)$ with a randomized estimator (essentially, adding a small random perturbation into the likelihood maximization problem). Adding noise is beneficial for computational reasons, since the set of $x \in \mathcal{X}$ whose MLE is exactly equal to $\widehat{\theta}_{\text{MLE}}(X)$ may be a challenging set to sample from. For certain examples, adding noise can also be beneficial from the statistical point of view, as for, e.g., the logistic regression setting, described in Model Class 1, where conditioning on the exact

MLE, $\widehat{\theta}_{\mathrm{MLE}}(X)$, may lead to a zero-power scenario. (We will discuss the role of σ further in Section 3.3 below.) In addition, we will also allow adding a twice-differentiable regularization function $\mathcal{R}(\theta)$ to the likelihood maximization problem, for instance $\mathcal{R}(\theta) \propto \|\theta\|^2$ for ridge regression, which may be beneficial in some applications.

Informally, our proposed aCSS algorithm takes the following form:

$$\begin{cases} X \sim P_{\theta_0}, \\ W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d), \\ \widehat{\theta} = \widehat{\theta}(X, W) = \arg\min_{\theta\in\Theta} \left\{ -\log f(X;\theta) + \mathcal{R}(\theta) + \sigma W^\top \theta \right\}, \\ \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \mid X, \widehat{\theta} \overset{\text{iid}}{\sim} P_{\widehat{\theta}}(\cdot \mid \widehat{\theta}), \end{cases}$$

where again $P_{\theta_0}(\cdot \mid \widehat{\theta})$ denotes the conditional distribution of $X \mid \widehat{\theta}$ when the data is distributed as $X \sim P_{\theta_0}$, and $P_{\widehat{\theta}}(\cdot \mid \widehat{\theta})$ is a plug-in estimate.

However, in many settings the penalized negative log-likelihood may not be strongly convex, or might even be nonconvex, in which case we will need to modify this procedure—while it is the case that, in many statistical problems, many tools exist that are likely to find the (perturbed) MLE with high probability (e.g., by carefully choosing a good initialization point), we will need to account for the fact that finding the global optimum is not guaranteed. Furthermore, in order to construct the copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$, we are implicitly assuming that we are able to generate i.i.d. samples from the conditional distribution of $X \mid \widehat{\theta}$. In practice, sampling directly from this density may be impossible, so we may need to turn to techniques such as Markov Chain Monte Carlo (MCMC), which can introduce dependence between the samples. Our next task, then, is to develop a more general and rigorous form of this simple algorithm, so that we can provide a practical method that can be deployed in a broad range of settings.

2.2 Algorithm for approximate co-sufficient sampling

In this section, we will formally define our aCSS algorithm. Below, we define our noisy estimator $\widehat{\theta}$ (Section 2.2.1), calculate the conditional distribution of $X \mid \widehat{\theta}$ (Section 2.2.2), and describe how to sample the copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ from the estimated conditional distribution (Section 2.2.3).

2.2.1 Sampling the estimator

Recall that the estimator $\widehat{\theta}$ is intended to be approximately equal to the MLE, even though it includes a regularization function and a random perturbation into the likelihood maximization problem. Writing

$$\mathcal{L}(\theta; x) = -\log f(x;\theta) + \mathcal{R}(\theta),$$

consider the optimization problem

$$\arg\min_{\theta\in\Theta} \mathcal{L}(\theta; X, W) \quad \text{where} \quad \mathcal{L}(\theta; x, w) = \mathcal{L}(\theta; x) + \sigma w^\top \theta, \tag{2.2}$$

where $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ is independent Gaussian noise, $\sigma > 0$ determines the noise level of the random perturbation, and $\mathcal{R} : \Theta \to \mathbb{R}$ is a twice-differentiable regularization function. In order to accommodate the penalized and unpenalized estimators with a single unified presentation, we can view the unpenalized version as a special case by simply taking $\mathcal{R}(\theta) \equiv 0$. (This type of randomly perturbed log-likelihood was previously studied by Tian and Taylor [2018], with the different aim of enabling selective inference on a high-dimensional parameter θ. In their work, the object of interest is the distribution of $\widehat{\theta}$, to enable inference on θ_0, whereas in our setting θ_0 is essentially a nuisance parameter.)

In the general setting where the negative log-likelihood might be nonconvex, the optimization problem (2.2) may be challenging—in particular, in the presence of nonconvexity, how would we find a global minimizer, and is a global minimizer even guaranteed to exist? In many settings, any available algorithm would only be able to guarantee that we find a first-order stationary point to (2.2) (if it even converges at all). To address this, we modify our procedure to allow $\widehat{\theta}$ to only *usually* be a well-behaved *local* optimum of (2.2). This enables aCSS to draw on the vast literature on optimizing penalized maximum likelihoods. Although the random perturbation by W makes (2.2) slightly non-standard for penalized maximum likelihood, the perturbation is linear in θ and hence has no impact on Hessians or convexity and only adds a fixed, trivially-computable constant vector to the gradient. Thus, although large linear perturbations can "tip over" an otherwise well-behaved basin of attraction, our theory will ensure this never happens asymptotically and in practice one can always choose σ small enough to make this astronomically unlikely; see Appendix D.1 for a more detailed discussion. In summary, we expect that any algorithm that empirically-often (it need not be provably-often) finds a local optimum for the unperturbed penalized maximum likelihood problem will suffice with almost no modification to solve (2.2) for the purposes required by the theory in this paper.

In particular, we will define $\widehat{\theta}$ to be any measurable function mapping a (data, noise) pair (x, w) to an estimate, i.e.,

$$\widehat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta,$$

and we will later assume this map is likely to return a strict second-order stationary point (SSOSP) of the minimization problem (2.2). Here we say that θ is a SSOSP of $\mathcal{L}(\theta; x, w)$ if two conditions are satisfied:

- θ is a first-order stationary point (FOSP) of $\mathcal{L}(\theta; x, w)$, meaning that $\nabla_\theta \mathcal{L}(\theta; x, w) = 0$ or equivalently $w = -\frac{\nabla_\theta \mathcal{L}(\theta; x)}{\sigma}$.
- The objective function is strictly convex at θ, i.e., $\nabla^2_\theta \mathcal{L}(\theta; x, w) \succ 0$ or equivalently $\nabla^2_\theta \mathcal{L}(\theta; x) \succ 0$.

We should think of $\widehat{\theta}(x, w)$ as the output of some optimization algorithm, such as gradient descent, being run to convergence on the minimization problem (2.2).

From this point on, abusing notation, depending on context we may write $\widehat{\theta}$ to denote the map $\widehat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta$, or may also write $\widehat{\theta}$ to denote $\widehat{\theta}(X, W)$, the random variable obtained by applying this map to the data.

2.2.2 Calculating the distribution conditioned on the estimator

Our next step is to calculate the conditional distribution of $X \,|\, \widehat{\theta}$, where $\widehat{\theta} = \widehat{\theta}(X, W)$ for random Gaussian noise $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$. As it turns out, it is generally not possible to do this exactly—in the rare degenerate case where $\widehat{\theta}(X, W)$ may fail to find a SSOSP of the optimization problem (2.2), we do not know the distribution of $\widehat{\theta} \,|\, X$ and therefore cannot calculate the distribution of $X \,|\, \widehat{\theta}$. We will avoid this degeneracy by conditioning on the event that $\widehat{\theta}(X, W)$ returns a SSOSP.

First, we assume some standard conditions on the parametric family, and a differentiability condition on the model and the regularization function (we will also assume implicitly that all the functions defined so far, namely, $\widehat{\theta}$, p, $\mathcal{L}$ and its derivatives, are measurable with respect to $\nu_{\mathcal{X}} \times \text{Leb}$ or $\nu_{\mathcal{X}}$ or Leb, as appropriate):

Assumption 1 (Regularity conditions)**.** *The family $\{P_\theta : \theta \in \Theta\}$ and regularization function $\mathcal{R}(\theta)$ satisfy:*

- *$\Theta \subseteq \mathbb{R}^d$ is a convex and open subset;*
- *For each $\theta \in \Theta$, P_θ has density $f(x;\theta) > 0$ with respect to the base measure $\nu_{\mathcal{X}}$;*
- *For each $x \in \mathcal{X}$, the function $\theta \mapsto \mathcal{L}(\theta; x) = -\log f(x;\theta) + \mathcal{R}(\theta)$ is continuously twice differentiable.*

We are now ready to calculate the conditional distribution of $X \,|\, \widehat{\theta}$.

Lemma 1. *Suppose Assumption 1 holds. Fix any $\theta_0 \in \Theta$, and let $(X, \widehat{\theta})$ be drawn from the joint model*

$$\begin{cases} X \sim P_{\theta_0}, \\ W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d), \\ \widehat{\theta} = \widehat{\theta}(X, W). \end{cases} \tag{2.3}$$

Suppose the event that $\widehat{\theta}$ is a SSOSP of $\mathcal{L}(\theta; X, W)$ has positive probability.

Then, conditional on this event, the conditional distribution of $X \,|\, \widehat{\theta}$ has density

$$p_{\theta_0}(\cdot \,|\, \widehat{\theta}) \propto f(x;\theta_0) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|^2}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}} \tag{2.4}$$

with respect to the base measure $\nu_{\mathcal{X}}$, where

$$\mathcal{X}_\theta = \left\{x \in \mathcal{X} : \text{ for some } w \in \mathbb{R}^d,\ \theta = \widehat{\theta}(x, w) \text{ is a SSOSP of } \mathcal{L}(\theta; x, w)\right\}. \tag{2.5}$$

The proof of this lemma is given in Appendix A.2. For intuition, we can consider the terms appearing in the calculation (2.4): the first term $f(x;\theta_0)$ expresses the original distribution of X (before conditioning), the second term $\exp\{\dots\}$ comes from the density of the multivariate normal distribution of W, the third term $\det(\dots)$ arises from a change-of-variables calculation when we move from the joint distribution of (X, W) to that of $(X, \widehat{\theta})$, and the

final term $\mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$ handles potential technical issues such as failure to find a SSOSP. In particular, the form of the second term is due to our choice of the multivariate normal distribution for the noise W; if we instead chose a different noise distribution, the results of this lemma would still hold if we make the appropriate changes to this second term (and the method would yield the same types of theoretical results as long as the distribution of W is continuous, supported everywhere on $\mathbb{R}^d$, and has similar concentration properties for $\|W\|$). In this work, we choose a multivariate normal distribution since the outcome of the procedure will therefore be invariant to rotations of the parameter space Θ; in settings where the choice of the basis for Θ is meaningful (e.g., we expect sparsity), it may be interesting to instead consider a non-rotationally-invariant noise distribution.

2.2.3 Sampling the copies

We next need to specify how to sample the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$. Below we describe several different approaches—which one we use will depend on the computational complexity of the problem at hand.

The i.i.d. sampling case In order to construct copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ that are exchangeable with the data X, we would like to sample the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ i.i.d. from the density $p_{\theta_0}(\cdot \,|\, \widehat{\theta})$, which by Lemma 1 specifies the exact conditional distribution of $X \,|\, \widehat{\theta}$. Since θ_0 is unknown we will use $\widehat{\theta}$ as a plug-in estimator. Our procedure is the following: after observing the data X,

$$\begin{cases} \text{Draw } W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d) \text{ and define } \widehat{\theta} = \widehat{\theta}(X, W). \\ \text{If } \widehat{\theta} \text{ is a SSOSP of } \mathcal{L}(\theta; X, W), \text{ then draw } \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \overset{\text{iid}}{\sim} p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}), \\ \qquad\qquad\qquad\qquad \text{otherwise return } \widetilde{X}^{(1)} = \ldots \widetilde{X}^{(M)} = X. \end{cases} \tag{2.6}$$

Here our estimated density for the conditional distribution of $X \,|\, \widehat{\theta}$ is given by

$$p_{\widehat{\theta}}(x \,|\, \widehat{\theta}) \propto f(x; \widehat{\theta}) \cdot \exp\left\{ -\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|^2}{2\sigma^2/d} \right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x) \right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}} \tag{2.7}$$

with respect to the base measure $\nu_{\mathcal{X}}$. (Lemma 4, in Appendix B.3, will verify that this expression indeed defines a valid density.)

Of course, in order to implement the sampling algorithm given in (2.6), we are implicitly assuming that it is computationally feasible to generate i.i.d. samples from $p_{\widehat{\theta}}(\cdot|\widehat{\theta})$. To avoid making this assumption, we next consider a more general framework.

The MCMC sampling case In the general case where sampling directly from $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$ may not be possible, we can instead use MCMC or any other strategy that ensures exchangeability. To be concrete, we will consider two schemes from Besag and Clifford [1989] for constructing the copies with MCMC sampling. Given $\widehat{\theta}$, let $\Pi(\cdot; x)$ be any collection of transition distributions, such that the density $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$ defines a stationary distribution. Assume that Π defines a reversible Markov chain. Given Π, we define two different schemes for generating the copies. (See Figure 2 for an illustration of these schemes.)

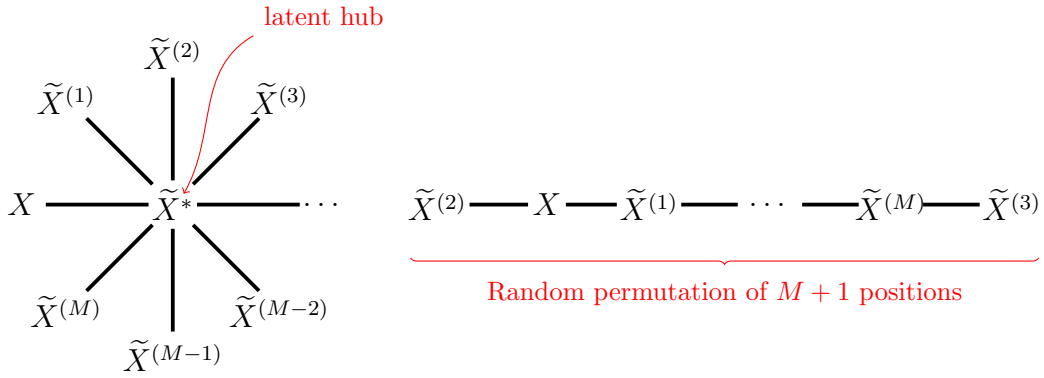

Figure 2: Left: the hub-and-spoke sampler. Right: the permuted serial sampler. In both diagrams, each thick black line represents running the reversible Markov chain for L steps.

- **Hub-and-spoke sampler**. Given X and $\widehat{\theta}$, we sample the copies as follows:
  - Initialize at X, and run the Markov chain for L steps to define the "hub" $\widetilde{X}^*$.
  - Independently for $m = 1, \ldots, M$, initialize at $\widetilde{X}^*$ and run the Markov chain for L steps to define the "spoke" $\widetilde{X}^{(m)}$.
- **Permuted serial sampler**. Given X and $\widehat{\theta}$, we sample the copies as follows:
  - Draw a uniform permutation π on $\{0, \ldots, M\}$ and find $m^* \in \{0, \ldots, M\}$ such that $\pi(m^*) = 0$.
  - Initialize at X, and run the Markov chain for Lm^* steps, stopping every L-th step to define the copies $\widetilde{X}^{(\pi(m^*-1))}, \ldots, \widetilde{X}^{(\pi(0))}$.
  - Independently, initialize at X, and run the Markov chain for $L(M-m^*)$ steps, stopping every L-th step to define the copies $\widetilde{X}^{(\pi(m^*+1))}, \ldots, \widetilde{X}^{(\pi(M))}$.

Later on, we will give concrete examples of how to implement these sampling schemes for specific models.

A unified definition To generalize our various options (i.i.d. sampling, hub-and-spoke MCMC sampling, and permuted serial MCMC sampling), we will write $\widetilde{P}_M(\cdot; X, \widehat{\theta})$ to denote the distribution of the collection of copies $(\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ conditional on X and $\widehat{\theta}$. For all three cases, our aCSS procedure for sampling the copies is the following:

$$\begin{cases} \text{Draw } W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d) \text{ and define } \widehat{\theta} = \widehat{\theta}(X, W). \\ \text{If } \widehat{\theta} \text{ is a SSOSP of } \mathcal{L}(\theta; X, W), \text{ then draw } (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \sim \widetilde{P}_M(\cdot; X, \widehat{\theta}), \\ \qquad\qquad\qquad\qquad \text{otherwise return } \widetilde{X}^{(1)} = \ldots \widetilde{X}^{(M)} = X. \end{cases} \tag{2.8}$$

In the i.i.d. sampling case, $\widetilde{P}_M(\cdot; X, \widehat{\theta})$ is simply equal to sampling from the product density $p_{\widehat{\theta}}(\cdot|\widehat{\theta}) \times \cdots \times p_{\widehat{\theta}}(\cdot|\widehat{\theta})$, and therefore depends on $\widehat{\theta}$ but not on X, while for the two MCMC samplers, there is dependence between the data and the copies even after conditioning on

$\widehat{\theta}$ (although, if the chain length L is sufficiently long, we would expect this dependence to be weak). Despite this dependence, all three of these sampling schemes satisfy the following exchangeability condition: for all $\theta \in \Theta$ with $\nu_{\mathcal{X}}(\mathcal{X}_\theta) > 0$,

$$\begin{aligned} &\text{If } X \sim p_\theta(\cdot \,|\, \theta) \text{ and } (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \,|\, X \sim \widetilde{P}_M(\cdot; X, \theta), \text{ then} \\ &\text{the random vector } (X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \text{ is exchangeable.} \end{aligned} \tag{2.9}$$

Note that $\widetilde{P}_M(\cdot; X, \theta)$ replaces *all* instances of $\widehat{\theta}$ in the definition of $\widetilde{P}_M(\cdot; X, \widehat{\theta})$ with θ's. Of course, it may be of interest to examine other sampling schemes, aside from the three described above. Our theoretical results below apply to any algorithm of the form (2.8) as long as the distribution $\widetilde{P}_M$ for drawing the copies is chosen to satisfy (2.9).

3 Theoretical results

In this section, we present our main result, proving a bound on the excess Type I error of any aCSS testing procedure.

3.1 Main result: Type I error bound

Before presenting the theorem, we will need a few more assumptions on the model and on the noisy estimator $\widehat{\theta}$. First, we need to assume that $\widehat{\theta}$ is (typically) an accurate estimator of the unknown true θ_0, and that $\widehat{\theta}$ will (typically) return a SSOSP for the optimization problem (2.2):

Assumption 2. *For any $\theta_0 \in \Theta$, the estimator $\widehat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta$ satisfies*

$$\mathbb{P}\left(\|\widehat{\theta}(X, W) - \theta_0\| \le r(\theta_0), \text{ and } \widehat{\theta}(X, W) \text{ is a SSOSP of } \mathcal{L}(\theta; X, W) \right) \ge 1 - \delta(\theta_0), \tag{3.1}$$

where the probability is taken with respect to the distribution $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$.

For many parametric families, the maximum likelihood estimator (or a penalized MLE) is typically shown to satisfy this type of condition with $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$ (here $\widetilde{\mathcal{O}}(\cdot)$ denotes that the scaling holds up to powers of $\log n$). This assumption has essentially the same flavor, except that our estimator $\widehat{\theta}$ is a random perturbation of the penalized MLE. We discuss this assumption in more detail in Appendix C.

Next, we place some assumptions on the derivatives of the log-likelihood. Let $H(\theta; x) = -\nabla^2_\theta \log f(x; \theta)$ and let $H(\theta) = \mathbb{E}_{\theta_0}[H(\theta; X)]$ (in particular, $H(\theta_0)$ is the Fisher information).

Assumption 3. *For any $\theta_0 \in \Theta$, the expectation $H(\theta)$ exists for all $\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta$, and furthermore*

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot \left(\lambda_{\max}(H(\theta) - H(\theta; X)) \right)_+ \right] \le \varepsilon(\theta_0) \tag{3.2}$$

and

$$\mathbb{E}_{\theta_0}\left[\exp\left\{ \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot \left(\lambda_{\max}(H(\theta; X) - H(\theta)) \right)_+ \right\} \right] \le e^{\varepsilon(\theta_0)}. \tag{3.3}$$

Here $r(\theta_0)$ is the same constant as appears in Assumption 2 (which, as mentioned above, will scale as $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$ in many settings). To interpret our assumption, we note that assumptions of the form

$$\|H(\theta; X) - H(\theta)\| = \mathcal{O}_{\mathrm{P}}(n^{1/2})$$

are standard for establishing classical results such as asymptotic normality of the MLE; even with a bound as weak as $r(\theta_0) = \mathrm{o}(n^{-1/4})$, this type of assumption will immediately imply that the first bound (3.2) holds. However, this type of condition is not quite sufficient for the theoretical arguments we need to establish, and we instead need the condition (3.3), which implies the same rate of convergence but with stronger control of the tails.

With our assumptions in place, we state the main result, which bounds the distance to exchangeability—and therefore, the Type I error—of any aCSS procedure.

Theorem 1. *Suppose Assumptions 1, 2, and 3 all hold. After observing the data X, suppose we run the aCSS algorithm (2.8), where the distribution $\widetilde{P}_M$ is chosen to satisfy (2.9). Then, if $X \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$, the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are approximately exchangeable with X, satisfying*

$$\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq 3\sigma \cdot r(\theta_0) + \delta(\theta_0) + \varepsilon(\theta_0).$$

In particular, this implies that for any predefined test statistic $T : \mathcal{X} \to \mathbb{R}$ and rejection threshold $\alpha \in [0, 1]$, the p-value defined in (1.1) satisfies

$$\mathbb{P}\left(\mathrm{pval}_T(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq \alpha\right) \leq \alpha + 3\sigma \cdot r(\theta_0) + \delta(\theta_0) + \varepsilon(\theta_0).$$

The proof of this theorem is given in Appendix A.1.

3.2 The asymptotic view

The theoretical guarantee given in Theorem 1 is nonasymptotic, but it typically implies asymptotic control of the Type I error. In particular, in many standard settings where the observed data arises from an independent sample of size n, the terms $r(\theta_0)$, $\delta(\theta_0)$, and $\varepsilon(\theta_0)$ are all vanishing, and in particular we will expect to see $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$. Thus, if we choose noise level $\sigma \asymp n^a$ for some $a < \frac{1}{2}$, this will lead to asymptotic Type I error control, i.e., $\mathbb{P}(\mathrm{pval} \leq \alpha) = \alpha + \mathrm{o}(1)$.

Furthermore, the Type I error bound in Theorem 1 gives insight into the role of approximate (or asymptotic) sufficiency in the method— $\widehat{\theta}(X, W)$ is essentially a MLE (assuming $\sigma = \mathrm{o}(n^{1/2})$ as before)—this is because the size of the perturbation of the negative log-likelihood, $\|\nabla_\theta \mathcal{L}(\theta; X, W) - \nabla_\theta \mathcal{L}(\theta; X)\| = \sigma\|W\| = \mathrm{o}(n^{1/2})$, is vanishing relative to $\|\nabla_\theta \mathcal{L}(\theta_0; X)\| \asymp n^{1/2}$. Thus under standard assumptions, $\widehat{\theta}(X, W)$ is asymptotically efficient, and inherits the asymptotic sufficiency properties of the MLE. At a high level, this means that the distributions $p_{\theta_0}(\cdot \,|\, \widehat{\theta})$ of $X \,|\, \widehat{\theta}$ and $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$ of $\widetilde{X}^{(m)} \,|\, \widehat{\theta}$ are asymptotically equal (i.e., the total variation distance between them is vanishing), leading to asymptotic exchangeability between X and its copies, and consequently an asymptotic Type I error bound at the nominal level α as shown in Theorem 1.

3.3 Choosing σ

It may seem odd that we have advocated for $\sigma > 0$ and yet the Type 1 error bound in our main result gets *worse* as σ increases. Indeed, increasing σ will generally degrade the Type 1 error of aCSS testing due to the fact that, as σ is increased, the method moves farther from conditioning on a sufficient statistic. And in fact, taking the limit as $\sigma \to 0$ in Theorem 1 gives the tightest possible Type 1 error bound (only Assumption 2 depends on σ, and in general we would expect it to be even more plausible for smaller σ). In addition, as discussed in Section 2.2.1 and in Appendix D.1, increasing σ can decrease the probability of finding an SSOSP for the optimization (2.2), which will not negatively impact the Type 1 error, but will decrease the power of the test by increasing the probability of returning a p-value of 1. However, despite these two downsides, there are two critical reasons why it is advantageous, and arguably necessary, to take $\sigma > 0$, and this is why we allow for it in Theorem 1.

First, note that if we took $\sigma = 0$, aCSS would need to sample from a distribution supported on a level set of the MLE function of X. This level set is a low-dimensional (and hence measure-zero) subset of $\mathcal{X}$, and thus it is generally computationally intractable to sample from exactly. There is some work on sampling a random variable conditional on the value of a function of it (e.g., Diaconis et al. [2013]), but only in very limited settings. Thus in most applications of aCSS, we are not aware of a computationally tractable approach that does not take $\sigma > 0$. Once we accept that $\sigma > 0$ is computationally necessary, the choice of its value represents a power-computation trade-off within the MCMC samplers we propose in this paper. This trade-off is discussed more in Appendix D.2, but essentially as σ approaches zero, it will take increasingly many MCMC steps (and associated computation) for the sampler to move "away" from the original X towards conditional independence. The more the sampler can move "away" from X, the higher the power of aCSS testing will tend to be, since a small p-value is obtained exactly when X stands out among the sampled copies.

Second, for models in which the MLE is sufficient for θ as well as for the parameters in a higher-dimensional supermodel of $\{P_\theta : \theta \in \Theta\}$ (e.g., in the logistic regression example the MLE is equivalent to X itself and thus is sufficient for all the parameters in any model), taking $\sigma = 0$ would lead to a completely powerless test for all alternatives in that supermodel. Exactly how large σ needs to be to break this degeneracy will likely need to be worked out on a case-by-case basis, and we defer a general treatment to future work. However, we see in Section 4 that for the logistic regression setting, described in Model Class 1, we can easily achieve high power with a σ value that is still sufficiently small to have no visible impact on the Type 1 error.

4 Examples

To provide further insight into the generality and practicality of aCSS testing, we establish that the necessary assumptions hold for four specific models. Example 1 (generalized linear models with canonical parameters) is an example of a regression model containing data with associated covariates, as discussed in Model Class 1. Example 2 (the Behrens–Fisher problem) and Example 3 (a Gaussian spatial process) are both examples of curved exponential families, discussed earlier in Model Class 2. Example 4 (a multivariate t distribution) is a

heavy-tailed model, and is thus an instance of Model Class 3.

In each case, we will see that the assumptions of Theorem 1 are satisfied with $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$, and with vanishing $\delta(\theta_0)$ and $\varepsilon(\theta_0)$. In particular, choosing a noise level $\sigma \asymp n^a$ for any $a < \frac{1}{2}$ is sufficient to ensure that the Type I error is asymptotically bounded by the nominal level α. We will then show simulation results for each of the four examples in Section 4.5 below.

4.1 Canonical generalized linear models (GLMs)

Example 1. We begin with the setting of a generalized linear model (GLM) with canonical parameters. Consider a logistic regression model with covariates $Z_i \in \mathbb{R}^d$ associated with each $X_i \in \mathbb{R}$, so that

$$f(x;\theta) = \prod_{i=1}^{n} \left(\frac{e^{Z_i^\top \theta}}{1+e^{Z_i^\top \theta}}\right)^{x_i} \cdot \left(\frac{1}{1+e^{Z_i^\top \theta}}\right)^{1-x_i},$$

parametrized by $\theta \in \Theta = \mathbb{R}^d$. (We interpret $f(x;\theta)$ as a density with respect to the base measure $\nu_{\mathcal{X}}$ on $\mathcal{X} = \mathbb{R}^n$ that places mass 1 on each point $x \in \{0,1\}^n$.) We can rewrite this in the notation of a generalized linear model (GLM),

$$f(x;\theta) = \exp\left\{x^\top Z\theta - \sum_{i=1}^{n} \log(1+e^{Z_i^\top \theta})\right\},$$

where $Z \in \mathbb{R}^{n\times d}$ is the matrix with rows Z_i. As discussed above, for $X \sim P_\theta$, the random vector $S(X) = Z^\top X \in \mathbb{R}^d$ provides a sufficient statistic; however, if the rows Z_i are in general position, then $Z^\top X$ will determine $X \in \{0,1\}^n$ uniquely, meaning that X is no longer random after we condition on $S(X) = Z^\top X$. In other words, co-sufficient sampling (CSS) would lead to zero power, and we therefore need to turn to aCSS testing.

More generally, we can consider any canonical GLM, of the form

$$f(x;\theta) = \exp\left\{x^\top Z\theta - \sum_{i=1}^{n} a(Z_i^\top \theta)\right\},$$

with respect to some base measure $\nu_{\mathcal{X}} = \mu \times \cdots \times \mu$ on $\mathcal{X} = \mathbb{R}^d$, where μ is a measure on $\mathbb{R}$. The function a is known as the *partition function*, and is strictly convex on its domain, which must be an open subset of $\mathbb{R}$. As for logistic regression, $Z^\top X$ is a sufficient statistic for $X \sim P_\theta$, but in the case of a discrete distribution (e.g., Poisson), CSS will again lead to zero power and so we should instead consider aCSS.

Suppose that the sample size n tends to infinity, while the parameter θ_0 is held constant (in particular, this implies that dimension d is held constant—we leave the high-dimensional setting for future work). For this example, and all the others below, we will consider the unpenalized version of the method, i.e., $\mathcal{R}(\theta) \equiv 0$. Assume the covariates are entrywise bounded, i.e., $\max_{i,j} \|Z_{ij}\|_\infty$ is bounded by a constant, and $\frac{1}{n}Z^\top Z \succeq \lambda_0 \mathbf{I}_d$ for a positive constant λ_0. We treat the covariates as fixed (i.e., the theory holds conditional on the covariates). Then, as we will show in Appendix C, for an appropriately-chosen initial estimator this example satisfies Assumptions 1, 2, and 3 with $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$, $\delta(\theta_0) = \mathcal{O}(n^{-1})$, and $\varepsilon(\theta_0) = 0$.

4.2 The Behrens–Fisher problem

Example 2. Next we consider the classical example of the Behrens–Fisher problem. Consider data

$$X_1^{(0)}, \dots, X_{n^{(0)}}^{(0)} \overset{\text{iid}}{\sim} \mathcal{N}(\mu^{(0)}, \gamma^{(0)}), \quad X_1^{(1)}, \dots, X_{n^{(1)}}^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}(\mu^{(1)}, \gamma^{(1)}),$$

with the two samples drawn independently. We are interested in testing the null hypothesis $H_0 : \mu^{(0)} = \mu^{(1)}$, and therefore the family of distributions can be parameterized by $\theta = (\mu, \gamma^{(0)}, \gamma^{(1)}) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \subseteq \mathbb{R}^3$, yielding a family $\{P_\theta : \theta \in \Theta\}$ where P_θ has density

$$f(x;\theta) = f(x;(\mu,\gamma^{(0)},\gamma^{(1)})) = \prod_{i=1}^{n^{(0)}} \frac{1}{\sqrt{2\pi\gamma^{(0)}}} e^{-(X_i^{(0)}-\mu)^2/2\gamma^{(0)}} \cdot \prod_{i=1}^{n^{(1)}} \frac{1}{\sqrt{2\pi\gamma^{(1)}}} e^{-(X_i^{(1)}-\mu)^2/2\gamma^{(1)}}$$

with respect to the Lebesgue measure on $\mathcal{X} = \mathbb{R}^{n^{(0)}+n^{(1)}}$.

This problem is an example of a curved exponential family (Problem Domain 2), for which the larger model is parametrized by $(\mu^{(0)}, \gamma^{(0)}, \mu^{(1)}, \gamma^{(1)})$—note that the constraint $\mu^{(0)} = \mu^{(1)}$ is a nonlinear constraint once we transform to the canonical parameters, which are given by $(\gamma^{(\ell)})^{-1}\mu^{(\ell)}$, $(\gamma^{(\ell)})^{-1}$ for each $\ell \in \{0, 1\}$. For this problem, under the null model (i.e., parametrized by $\theta = (\mu, \gamma^{(0)}, \gamma^{(1)})$), the minimal sufficient statistic is nonetheless four-dimensional—for example, the sample means and sample standard deviations of $\{X_i^{(0)}\}$ and of $\{X_i^{(1)}\}$ form a minimal sufficient statistic. Of course, this statistic is also sufficient for the larger alternative model (where $\mu^{(0)} \neq \mu^{(1)}$); once we condition on this sufficient statistic, the remaining randomness in the data carries no information about the parameters $\mu^{(0)}$ and $\mu^{(1)}$. Therefore, CSS would lead to a completely powerless procedure, and we instead turn to aCSS. (As mentioned earlier in Section 1.3, Lillegård [2001] mention the possibility of, but do not pursue, an aCSS-like procedure for this specific example.)

Suppose that the sample size n tends to infinity, while the parameter θ_0 is held constant and the ratio $\frac{\max\{n^{(0)},n^{(1)}\}}{\min\{n^{(0)},n^{(1)}\}}$ is bounded by a constant. Then, as we will show in Appendix C, for an appropriately-chosen initial estimator this example satisfies Assumptions 1, 2, and 3 with $r(\theta_0) \asymp \widetilde{\mathcal{O}}(n^{-1/2})$, $\delta(\theta_0) \asymp \mathcal{O}(n^{-1})$, and $\varepsilon(\theta_0) = \widetilde{\mathcal{O}}(n^{-1})$.

4.3 A Gaussian spatial process

Example 3. For our next example, we will work in a dependent data setting—unlike the other three examples, we do not have independent observations. Our model is a Gaussian spatial process. Suppose that $X \in \mathbb{R}^n$ is distributed according to a multivariate Gaussian,

$$X \sim \mathcal{N}(0, \Sigma_\theta),$$

where the covariance matrix Σ_θ is parametrized by a scalar $\theta \in \mathbb{R}$. Specifically, we will consider a spatial Gaussian process where

$$(\Sigma_\theta)_{ij} = \exp\{-\theta \cdot D_{ij}\},$$

where $(D_{ij}) \in \mathbb{R}^{n\times n}$ is a pairwise distance matrix among n spatial points. In other words, we can think of the observation X_i as corresponding to a location $z_i \in \mathbb{R}^k$ for some ambient

dimension k, and the correlation between X_i and X_j is a decaying function of the distance between locations z_i and z_j, i.e., $D_{ij} = \|z_i - z_j\|$. We assume that the distances D_{ij} are known, and the parameter $\theta \in \Theta = (0, \infty) \subseteq \mathbb{R}$ is the only unknown. This example, like Example 2, is an instance of a curved exponential family. In this case, the larger model is given by $X \sim \mathcal{N}(0, \Omega^{-1})$, where the inverse covariance Ω is the canonical parameter. The nonlinear constraints introduced by the spatial model take the form

$$(D_{ij})^{-1} \log(\Omega^{-1})_{ij} = (D_{k\ell})^{-1} \log(\Omega^{-1})_{kl}$$

for all indices i, j, k, ℓ (since the expression on each side of this equation should be equal to the same value θ). As in Example 4, the minimal sufficient statistic for our curved exponential null model is the same as that for the larger exponential family—in this case, it is given by the (uncentered) sample covariance—and therefore CSS would result in a powerless procedure for testing against any mean-zero multivariate Gaussian alternative.

Now we turn to aCSS for this example. In this setting, the distribution P_θ has density

$$f(x; \theta) = \frac{1}{(2\pi)^{n/2} \det(\Sigma_\theta)^{1/2}} e^{-x^\top \Sigma_\theta^{-1} x/2},$$

with respect to the Lebesgue measure on $\mathbb{R}^n$. The negative log-likelihood $\theta \mapsto -\log f(x; \theta)$ is therefore nonconvex, due to the nature of the map $\theta \mapsto \Sigma_\theta$.

It is known, however, that in the special case where the locations z_i are on a regular integer lattice, standard results such as asymptotic normality of the MLE can be obtained [Bachoc, 2014], and so we will work in this setting. Consider the integer grid $\{z_1, \ldots, z_n\} = \{1, \ldots, N\}^k$, where $n = N^k$. As above, the distances D_{ij} are given by $\|z_i - z_j\|$. Suppose that the grid size N tends to infinity, while the dimension k and the parameter θ_0 are held constant. Then, as we will show in Appendix C, for an appropriately-chosen initial estimator this example satisfies Assumptions 1, 2, and 3 with $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$, $\delta(\theta_0) = \mathcal{O}(n^{-1})$, and $\varepsilon(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$.

4.4 The multivariate t distribution with unknown covariance

Example 4. Our last example will demonstrate that our methodology can be applied even in settings where the data is extremely heavy-tailed—specifically, the multivariate t distribution. We consider a setting with n i.i.d. draws from a zero-mean multivariate t distribution,

$$X_i \overset{\text{iid}}{\sim} t_\gamma(0, \theta^{-1}),$$

where $\theta^{-1} \in \mathbb{R}^{k \times k}$ is an unknown covariance matrix while $\gamma > 0$ is the known degrees-of-freedom parameter. (Breaking with standard notation, we will use a lowercase θ to denote a matrix parameter, to agree with our notation throughout this paper.) Our family of distributions is therefore given by $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^{k \times k}$ is the set of positive definite $k \times k$ matrices. We can view Θ as a convex open subset of $\mathbb{R}^d$ with $d = \frac{k(k+1)}{2}$, by considering the upper triangle of a positive definite matrix θ. The density is

$$f(x; \theta) = \prod_{i=1}^n c_{k,\gamma} \det(\theta)^{1/2} \left(\gamma + x_i^\top \theta x_i\right)^{-\frac{\gamma+k}{2}},$$

with respect to the Lebesgue measure on $\mathcal{X} = (\mathbb{R}^k)^n$, where $c_{k,\gamma}$ depends only on the dimension k and the degrees-of-freedom parameter γ, and not on the unknown parameter θ. Unlike a GLM, we cannot write the log-density $\log f(x;\theta)$ in the form (function of x)$\cdot$(function of θ). In fact, we can see that, up to permutation and/or multiplication by -1 of the data points $i = 1, \ldots, n$, the data X itself is a minimal sufficient statistic for θ, so there is no sufficient statistic that would not essentially fully specify the data. Thus for instance, CSS testing would be powerless against any i.i.d. alternative that is invariant to reflection through the origin. However, the approximate sufficiency framework is well-suited for this example.

Suppose that the sample size n tends to infinity, while the degrees-of-freedom parameter γ and the unknown matrix parameter θ_0 are held constant (in particular, this implies that the dimension k is held constant—we leave the high-dimensional setting for future work). Then, as we will show in Appendix C, for an appropriately-chosen initial estimator this example satisfies Assumptions 1, 2, and 3 with $r(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$, $\delta(\theta_0) = \mathcal{O}(n^{-1})$, and $\varepsilon(\theta_0) = \widetilde{\mathcal{O}}(n^{-1/2})$.

4.5 Simulations

We now demonstrate the performance of aCSS for each of the four examples described above; code to reproduce the simulations is available at `http://www.stat.uchicago.edu/~rina/code/aCSS.zip`. We will first show two examples in Section 4.5.1 with relatively simple parametric alternative models, for which competing methods exist; in these examples, we will see aCSS testing is as powerful as the most powerful established method, namely, the score test. Then, in Section 4.5.2, we will consider two more complex examples exhibiting alternative models which elude standard approaches, and for which we are unaware of any existing test that would be powerful; we will see that aCSS testing can be powerful in such settings through the choice of a relatively sophisticated test statistic that fully leverages the particular alternative model.

For both types of examples, we will also see that the aCSS test is empirically valid (the rejection probability is almost exactly the nominal level $\alpha = 0.05$ under the null hypothesis) and that it has only slightly less power than an oracle method—this oracle method is given extra information about the distribution that reduces the composite null to a simple null, and computes a p-value (1.1) by applying the same statistic function T as aCSS to M copies $\widetilde{X}^{(m)}$ drawn independently (unconditionally) from that simple null.

4.5.1 Simulations with a parametric alternative

We use Examples 2 (Behrens–Fisher) and 4 (multivariate t) to demonstrate similar power between the aCSS test and the score test under parametric alternatives. The results, plotted in Figure 3, show the aCSS tests have very similar power to both the oracle and score tests. The simulation setups for the two examples are summarized below; the choice of the proposal distributions for the MCMC samplers, and chain lengths L, are described in detail in Appendix D.

Example 2 (Behrens–Fisher) For the Behrens–Fisher example, the alternative model is as described in Section 4.2 but with $(\mu^{(0)}, \mu^{(1)}, \gamma^{(0)}, \gamma^{(1)})$ unconstrained in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$.

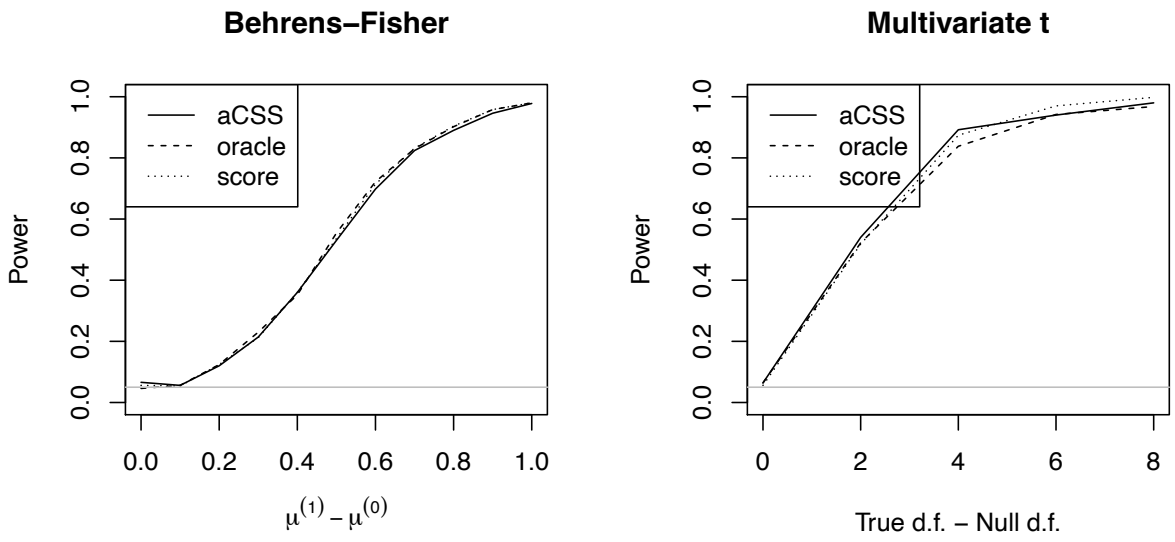

Figure 3: Power of the aCSS test compared to an unconditional oracle that knows the (simple) null hypothesis, and compared also to the score test, for the two examples discussed in Section 4.5.1. The aCSS test controls the Type I error at the nominal 5% level (dotted line) under the null (represented by 0 on the x-axis in each plot), and has very similar power to the oracle and score test under the alternatives. Each point represents 500 independent replications, with the maximum standard error $\approx 2\%$ and the standard error at the left edge of each plot (under the null) below 1%.

- To generate the data, we take $n^{(0)} = n^{(1)} = 50$, $\mu^{(0)} = 0$, $\gamma^{(0)} = 1$, $\gamma^{(1)} = 2$, and $\mu^{(1)} \in \{0, 0.1, 0.2, \ldots, 1\}$ (with $\mu^{(1)} = 0$ corresponding to the case where the null hypothesis holds).
- The test statistic T (used both for aCSS and for the oracle) is given by the absolute difference in sample means between the two halves of the data.
- aCSS is run with the hub-and-spoke sampler with parameters $\sigma^2 = 1$ and $M = 500$. The oracle method is given all parameter values except for $\mu^{(1)}$, so that the null $\mu^{(1)} = 0$ is simple.

Example 4 (multivariate t) For the multivariate t example, the alternative model is as described in Section 4.4 but with the degrees-of-freedom parameter γ unknown and unconstrained (aside from being positive).

- To generate the data, we let $n = 100$, $\theta_0 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 2 \end{pmatrix}$, and $\gamma = 2$ be the assumed degrees of freedom under the null hypothesis ("$\text{d.f.}_{\text{null}}$"). The distribution of the data is given by $t_{\text{d.f.}}(0, \theta_0^{-1})$, where the degrees of freedom "d.f." is taken from $\{2, 4, 6, 8, 10\}$. Therefore $\text{d.f.} = 2$ represents the case where the null is true, and $\text{d.f.} - \text{d.f.}_{\text{null}}$ measures the deviation from the null hypothesis.
- The test statistic T (used both for aCSS and for the oracle) is chosen to be the same as for the score test.
- aCSS is run with the hub-and-spoke sampler with parameters $\sigma^2 = 1$ and $M = 100$. The oracle method is given all parameter values except for γ, so that the null $\gamma = 2$ is simple.

4.5.2 Simulations without a parametric alternative

We use Examples 1 and 3 to demonstrate the power of aCSS testing under more complex alternative models for which no existing methods (including the score test) are suitable. The results, plotted in Figure 4, show the aCSS tests have very similar power to the oracle. For the four examples, the settings of the simulation are as follows. In each case, the choice of the proposal distribution for the MCMC sampler, and chain length L, are described in detail in Appendix D.

Example 1 (logistic regression) For the logistic regression example, we use aCSS to test a conditional independence hypothesis, so there is a response variable Y that, under the alternative, changes the conditional distribution of $X \mid Z$ given in Section 4.1. Y is drawn from a nonparametric model which is well approximated by a single index model, but does not exactly follow this model.

- To generate the data, we take $n = 100$, and $X \mid Z$ follows 5-dimensional logistic regression with coefficient vector $\theta_0 = 0.2 \cdot \mathbf{1}$. Y's conditional distribution is given by: $Y \mid (Z, X = 0) = f_0(g_0(Z) + \beta_0^\top Z) + \mathcal{N}(0, 1)$ and $Y \mid (Z, X = 1) = f_1(g_1(Z) + \beta_1^\top Z) +$

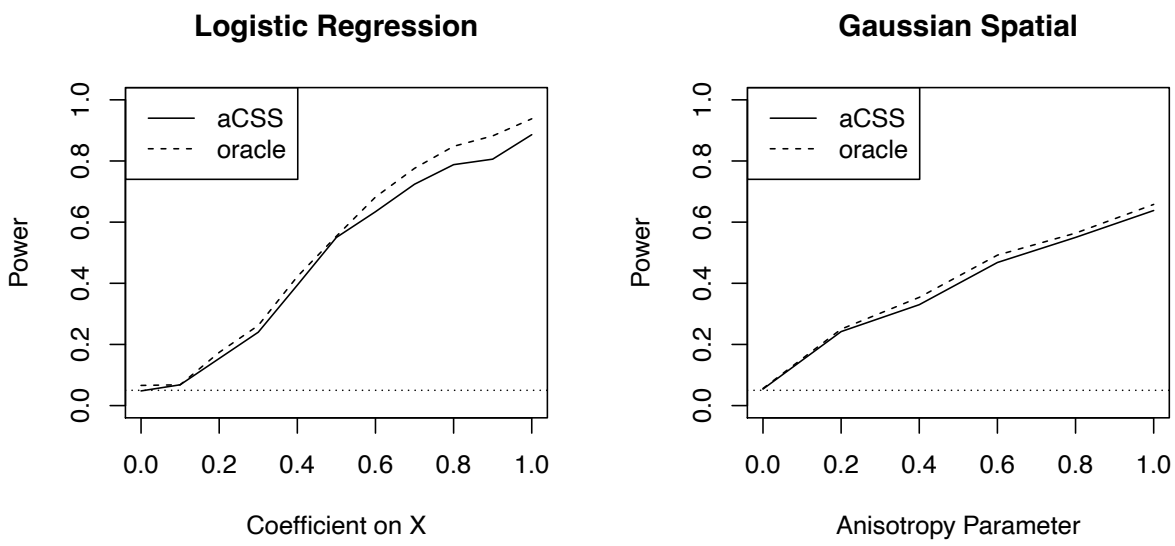

Figure 4: Power of the aCSS test compared to an unconditional oracle that knows the (simple) null hypothesis, for the two examples discussed in Section 4.5.1. The aCSS test controls the Type I error at the nominal 5% level (dotted line) under the null (represented by 0 on the x-axis in each plot), and has very similar power to the oracle and score test under the alternatives. Each point represents 500 independent replications, with the maximum standard error $\approx 2\%$ and the standard error at the left edge of each plot (under the null) below 1%.

$\mathcal{N}(0,1)$. We choose $f_0 = f_1 = t \mapsto t + 0.5t^3$, $g_0 = g_1 = z \mapsto 0.5\sum_{j=1}^{5}(z_j)_+$, $\beta_0 = c \cdot \mathbf{e}_1$, and $\beta_1 = c \cdot \mathbf{e}_5$, where $c \in \{0, 0.1, 0.2, \ldots, 1\}$ indicates the signal strength (with $c = 0$ corresponding to the null hypothesis). The nonlinearity of g_0 and g_1 means that the single index model does not exactly describe the conditional distribution of Y.

- The test statistic T (used both for aCSS and for the oracle) is computed by estimating the coefficient vector on Z in a single index model via sliced inverse regression [Li, 1991] separately on the data sets $\{(Y_i, Z_i) : X_i = 0\}$ and $\{(Y_i, Z_i) : X_i = 1\}$, respectively (though recall that the single index model does not strictly hold for either data set), and then computing the angle between these estimated coefficient vectors.

- aCSS is run with the hub-and-spoke sampler with parameters $\sigma^2 = 10$ and $M = 500$. To implement the oracle method in this example, the oracle is given the distribution of $X \mid Z$, i.e., the true coefficient vector θ_0 for the logistic regression model; under the null hypothesis, $X \mid Z, Y$ follows the same distribution as $X \mid Z$, and thus the oracle is given full knowledge of the distribution of $X \mid Z, Y$ under the null.

Example 3 (Gaussian spatial) For the Gaussian spatial process example, we take a 2-dimensional 10×10 integer lattice $\{1, \ldots, 10\}^2$ for the spatial points.

- The distribution of the data is as described in Example 3 with the exception that there exists a line $\mathcal{L}$ bisecting the lattice, and for two points i and j whose positions (z_i and z_j, respectively) are on opposite sides of $\mathcal{L}$, instead of their covariance being given by $e^{-\theta_0\|z_i - z_j\|}$, it is instead given by $(1-c)e^{-\theta_0\|z_i - z_j\|}$. For instance, the data points could come from soil samples, and $\mathcal{L}$ might be a possible geological ridge reducing the dependence between points on either side of it. In our experiments, $\theta_0 = 0.2$, $\mathcal{L}$ is horizontal with intercept 5.5 so that 50 of the lattice points lie below it and the other 50 lie above it, and $c \in \{0, 0.2, \ldots, 1\}$ is an anisotropy parameter, with $c = 0$ indicating an isotropic spatial process so that the null hypothesis holds.

- The test statistic T (used both for aCSS and for the oracle) is computed as follows. We first compute a thresholded kernel matrix $\Delta \in \mathbb{R}^{n \times n}$ with entries $\Delta_{i,j} = e^{-|X_i - X_j|}\mathbb{1}_{\|z_i - z_j\| = 1}$ and then use Δ as the kernel matrix for spectral clustering with two clusters. Denoting the two clusters as S and S^c, the value of T is then computed as the normalized negative sum of kernel distances between the two groups:

$$-\left(\frac{1}{\sum_{i \in S^c, j \in S \cup S^c} \Delta_{i,j}} + \frac{1}{\sum_{i \in S, j \in S \cup S^c} \Delta_{i,j}}\right) \sum_{i \in S, j \in S^c} \Delta_{i,j}.$$

- aCSS is run with the hub-and-spoke sampler with parameters $\sigma^2 = 1$ and $M = 100$. The oracle method is given θ_0, $\mathcal{L}$, and the functional form for Σ in terms of c, so that the null $c = 0$ is simple.

5 Discussion

Approximate co-sufficient sampling offers a new framework for inference on goodness-of-fit and related problems such as conditional independence testing and inference on target parameters, under mild assumptions on a composite null model. In this section, we will first revisit the construction of aCSS to develop a deeper intuition for the ideas behind the method, and will then examine some open questions and directions that remain.

5.1 The importance of conditioning: comparison to the parametric bootstrap

Here we return to the construction of the aCSS method, with new insights obtained from the proof of our main result, Theorem 1. In particular, why is it important to condition on $\widehat{\theta}$ when we sample the copies?

In the construction of aCSS, after conditioning on $\widehat{\theta}$, we sample copies $\widetilde{X}^{(m)}$ that are approximately exchangeable with X as long as it holds that $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}) \approx p_{\theta_0}(\cdot \,|\, \widehat{\theta})$. This is because, conditional on $\widehat{\theta}$, the copies are sampled from the density $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$, while the unknown true null density of $X \,|\, \widehat{\theta}$ is instead $p_{\theta_0}(\cdot \,|\, \widehat{\theta})$; we simply use $\widehat{\theta}$ as a plug-in estimator of θ_0 to define the distribution from which we sample the copies. In our proofs, we saw that aCSS leads to asymptotically valid tests as long as $\mathsf{d}_{\mathsf{TV}}(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}))$ is vanishing.

It is tempting to ask whether the same idea can be used without conditioning on $\widehat{\theta}$. That is, since the true data is distributed as $X \sim P_{\theta_0}$ under the null, can we plug in $\widehat{\theta}$ for θ_0 and sample the copies $\widetilde{X}^{(m)}$ from $P_{\widehat{\theta}}$? In fact, this non-conditional version of the procedure is simply recovering the parametric bootstrap—and, as we observed in Section 1, the parametric bootstrap may result in inflated Type I error rates in certain settings, depending on the test statistic T that we use. This is because, in general, it will not be the case that $\mathsf{d}_{\mathsf{TV}}(P_{\theta_0}, P_{\widehat{\theta}})$ is vanishing, even for $\widehat{\theta}$ chosen to be the MLE, and therefore, if we define the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ by sampling (unconditionally) from $P_{\widehat{\theta}}$, rather than from the conditional distribution estimate $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$, it will generally be the case that, for some adversarially chosen test statistic $T(X)$, we may have Type I error that exceeds the nominal level α by a nonvanishing amount.

5.2 Can we condition on less information?

More generally, what if we consider conditioning on a different statistic $S = S(X)$ (or a perturbed version $S = S(X, W)$), which contains strictly less information about the data X than the (perturbed) MLE $\widehat{\theta}$? Of course, the above unconditional distribution is simply the extreme case of this idea, since it conditions on no information at all. Can we choose S so that it reveals less information about X and thus yields potentially higher power against the alternative, while still retaining approximate validity of our test? To run such a test, we would need to sample the copies from the plug-in estimated distribution $P_{\widehat{\theta}}(\cdot \,|\, S)$ rather than the true conditional null distribution $P_{\theta_0}(\cdot \,|\, S)$ of $X \,|\, S$, and in order for the copies to be approximately exchangeable with X under the null, we will need this plug-in estimate to be accurate, i.e., $P_{\widehat{\theta}}(\cdot \,|\, S) \approx P_{\theta_0}(\cdot \,|\, S)$—in other words, S needs to be (approximately) sufficient.

As discussed earlier in Section 3.2, the perturbed MLE $\widehat{\theta}$ is asymptotically sufficient under standard conditions; since $\widehat{\theta}$ has the same dimension d as the true parameter θ_0, it is clear that it is also (asymptotically) a *minimal* sufficient statistic. Therefore, if we choose to condition on any other statistic S, if S contains strictly less information about the data X than $\widehat{\theta}$, the approximate validity of aCSS would no longer hold.

5.3 Open questions

Given our new framework for inference via approximate co-sufficient sampling, many open questions remain regarding the properties of this framework, and the settings in which it can be applied.

1. *Power.* How does the choice of statistic T interact with the aCSS framework, to offer the best possible power? In particular, might it be the case that the choice of T that is most powerful under an aCSS test is not the same as the T that is most powerful for an oracle test (with a known point null hypothesis, i.e., θ_0 known)?

2. *Computation.* Are there particular algorithms that enable efficient sampling of the copies $\widetilde{X}^{(m)}$, or are there statistics T that allow us to calculate $T(\widetilde{X}^{(m)})$ without needing to fully observe $\widetilde{X}^{(m)}$—for example, through leveraging symmetries in the model and the conditional distribution?

3. *Additional models.* In addition to the examples described in this paper, can the aCSS framework be applied to similar problems such as non-canonical generalized linear models, low-rank regression, or rank-based data? Moving to more challenging settings, does the aCSS framework extend to latent variable models, errors-in-variables models, or models with missing data?

4. *Broader settings.* Can aCSS be applied in a nonparametric setting (perhaps with constraints on the statistics T allowed)? Is aCSS robust to model misspecification?

5. *Relaxing regularity conditions and extending to high dimensions.* Can aCSS be applied in settings where the null model is d-dimensional, but cannot be represented as a convex and open subset of $\mathbb{R}^d$? For instance, we may have sparsity constraints (with the parameter space given by all s-sparse vectors in $\mathbb{R}^p$) or rank constraints (with the parameter space consisting of all matrices with rank at most r in $\mathbb{R}^{a\times b}$). It would appear that any extension of aCSS testing to high dimensions would require incorporating some such low-dimensional structure, in order to ensure the existence of a non-degenerate approximately sufficient statistic, as well as a consistent estimator $\widehat{\theta}$.

A Proofs of main results

Before presenting the proofs of our theoretical results, we first establish some notation that we will use throughout these proofs. Let

$$\Omega_{\text{SSOSP}} = \left\{(x,w) \in \mathcal{X} \times \mathbb{R}^d : \widehat{\theta}(x,w) \text{ is a SSOSP of } \mathcal{L}(\theta; x, w)\right\},$$

and let

$$\Psi_{\text{SSOSP}} = \{(x, \theta) \in \mathcal{X} \times \Theta : x \in \mathcal{X}_\theta\},$$

where $\mathcal{X}_\theta$ is defined as in (2.5). The following lemma (proved in Appendix B.1) establishes a bijection between these sets:

Lemma 2. *Under Assumption 1, the map*

$$\psi : (x, w) \mapsto (x, \widehat{\theta}(x, w))$$

defines a bijection between Ω_{SSOSP} *and* Ψ_{SSOSP}*, with inverse*

$$\psi^{-1} : (x, \theta) \mapsto \left(x, -\frac{\nabla_\theta \mathcal{L}(\theta; x)}{\sigma}\right).$$

A.1 Proof of Theorem 1

Define $P^*_{\theta_0}$ to be the distribution of $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditional on the event $(X, W) \in \Omega_{\text{SSOSP}}$. (If this event has probability 0 then the theorem holds trivially, so we can ignore this case.) Consider the joint distribution

$$\text{Distrib. (a):} \quad \begin{cases} (X, W) \sim P^*_{\theta_0}, \\ \widehat{\theta} = \widehat{\theta}(X, W), \\ \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \,|\, X, \widehat{\theta} \sim \widetilde{P}_M(\cdot; X, \widehat{\theta}), \end{cases}$$

which is clearly equivalent to the aCSS procedure (2.8) if we condition on the event $(X, W) \in \Omega_{\text{SSOSP}}$. On the other hand, on the event that $(X, W) \notin \Omega_{\text{SSOSP}}$, then by definition we set $\widetilde{X}^{(1)} = \cdots = \widetilde{X}^{(M)} = X$, and so exchangeability can only be violated on the event Ω_{SSOSP}. Therefore, we have

$$\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \le \mathsf{d}_{\mathsf{exch}}\big(\text{Distribution of } X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \text{ under Distrib. (a)}\big). \tag{A.1}$$

(We formalize this intuition in Lemma 3 in Appendix B.2.)

Next, let $Q^*_{\theta_0}$ be the marginal distribution of $\widehat{\theta}(X, W)$ under $(X, W) \sim P^*_{\theta_0}$, and define

$$\text{Distrib. (b):} \quad \begin{cases} \widehat{\theta} \sim Q^*_{\theta_0}, \\ X \,|\, \widehat{\theta} \sim p_{\theta_0}(\cdot \,|\, \widehat{\theta}), \\ \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \,|\, X, \widehat{\theta} \sim \widetilde{P}_M(\cdot; X, \widehat{\theta}). \end{cases}$$

where $p_{\theta_0}(\cdot \,|\, \widehat{\theta})$ is defined as in Lemma 1. By definition of $Q^*_{\theta_0}$, together with Lemma 1, we can see that the joint distribution of $(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)})$ under Distrib. (b), is equal to its joint distribution under Distrib. (a), and therefore

$$\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \le \mathsf{d}_{\mathsf{exch}}\big(\text{Distribution of } X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \text{ under Distrib. (b)}\big).$$

Finally, we define another distribution,

$$\text{Distrib. (c):} \qquad \begin{cases} \widehat{\theta} \sim Q^*_{\theta_0}, \\ X \,|\, \widehat{\theta} \sim p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta}), \\ \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \,|\, X, \widehat{\theta} \sim \widetilde{P}_M(\cdot; X, \widehat{\theta}). \end{cases}$$

(As mentioned earlier, Lemma 4 in Appendix B.3 will verify that the density $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$ exists almost surely over $\widehat{\theta}$.) Since $\widetilde{P}_M(\cdot; X, \theta)$ was constructed to satisfy (2.9), it holds that under Distrib. (c), the random variables $(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ are exchangeable (in fact, they are exchangeable conditional on $\widehat{\theta}$). Therefore, by definition of $\mathsf{d}_{\mathsf{exch}}$, we have

$$\mathsf{d}_{\mathsf{exch}}\big(\text{Distribution of } X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \text{ under Distrib. (b)}\big) \le \mathsf{d}_{\mathsf{TV}}\big(\text{Distrib. (b)}, \text{Distrib. (c)}\big),$$

and comparing the definitions of Distrib. (b) and Distrib. (c), it is easy to verify that

$$\mathsf{d}_{\mathsf{TV}}(\text{Distrib. (b)}, \text{Distrib. (c)}) = \mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right].$$

Combining everything, we have shown that the aCSS procedure (2.8) satisfies

$$\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \le \mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right]. \tag{A.2}$$

We next need to bound this expected total variation.

We begin with the well known expression for total variation distance between two densities g, h, which is given by $\mathsf{d}_{\mathsf{TV}}(g, h) = \mathbb{E}_g\left[\left(1 - \frac{h(X)}{g(X)}\right)_+\right]$. Therefore,

$$\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] = \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[\left(1 - \frac{p_{\widehat{\theta}}(X \,|\, \widehat{\theta})}{p_{\theta_0}(X \,|\, \widehat{\theta})}\right)_+\right]\right]. \tag{A.3}$$

Recalling the definitions (2.4) and (2.7) (and noting in particular that these two densities have the same support by definition), after calculating normalizing constants we can verify that

$$\frac{p_{\widehat{\theta}}(x \,|\, \widehat{\theta})}{p_{\theta_0}(x \,|\, \widehat{\theta})} = \frac{\frac{f(x;\widehat{\theta})}{f(x;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right]}. \tag{A.4}$$

Next we take a Taylor series for the function $\theta \mapsto \log f(X; \theta)$. For any x, θ we can calculate

$$\log\left(\frac{f(x;\theta_0)}{f(x;\theta)}\right) = (\theta_0 - \theta)^\top \nabla_\theta \log f(x;\theta) + \int_{t=0}^1 (1-t) \cdot (\theta_0 - \theta)^\top \nabla^2_\theta \log f(x;\theta_t)(\theta_0 - \theta)\, \mathsf{d}t,$$

where we write $\theta_t = (1-t)\theta_0 + t\theta$. Therefore, for any x, x' we have

$$\begin{aligned}
\frac{\frac{f(x';\theta)}{f(x';\theta_0)}}{\frac{f(x;\theta)}{f(x;\theta_0)}} &= \exp\left\{\log\left(\frac{f(x;\theta_0)}{f(x;\theta)}\right) - \log\left(\frac{f(x';\theta_0)}{f(x';\theta)}\right)\right\} \\
&= \exp\Bigg\{ -(\theta_0-\theta)^\top\left(\nabla_\theta \log f(x';\theta) - \nabla_\theta \log f(x;\theta)\right) \\
&\qquad - \int_{t=0}^1 (1-t)\cdot(\theta_0-\theta)^\top\left(\nabla_\theta^2 \log f(x';\theta_t) - \nabla_\theta^2 \log f(x;\theta_t)\right)(\theta_0-\theta)\,\mathsf{d}t\Bigg\} \\
&= \exp\Bigg\{(\theta_0-\theta)^\top\left(\nabla_\theta \mathcal{L}(\theta;x') - \nabla_\theta \mathcal{L}(\theta;x)\right) \\
&\qquad + \int_{t=0}^1 (1-t)\cdot(\theta_0-\theta)^\top\left(H(\theta_t;x') - H(\theta_t;x)\right)(\theta_0-\theta)\,\mathsf{d}t\Bigg\} \\
&\le \exp\Bigg\{(\theta_0-\theta)^\top\left(\nabla_\theta \mathcal{L}(\theta;x') - \nabla_\theta \mathcal{L}(\theta;x)\right) \\
&\qquad + \frac{1}{2}\sup_{t\in[0,1]}(\theta_0-\theta)^\top\left(H(\theta_t;x') - H(\theta_t;x)\right)(\theta_0-\theta)\Bigg\},
\end{aligned}$$

where the inequality holds since $\int_{t=0}^1 (1-t)\cdot h(t)\ \mathsf{d}t \le \int_{t=0}^1 (1-t)\ \mathsf{d}t \cdot \sup_{t\in[0,1]} h(t) = \frac{1}{2}\sup_{t\in[0,1]} h(t)$ for any function $h:\mathbb{R}\to\mathbb{R}$. For any $\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta$, it therefore holds that, for all x, x',

$$\begin{aligned}
\frac{\frac{f(x';\theta)}{f(x';\theta_0)}}{\frac{f(x;\theta)}{f(x;\theta_0)}} &\le \exp\Bigg\{r(\theta_0)\left(\|\nabla_\theta \mathcal{L}(\theta;x')\| + \|\nabla_\theta \mathcal{L}(\theta;x)\|\right) \\
&\qquad + \frac{r(\theta_0)^2}{2}\sup_{\theta'\in\mathbb{B}(\theta_0,r(\theta_0))\cap\Theta}\lambda_{\max}\left(H(\theta';x') - H(\theta';x)\right)\Bigg\} \le \exp\left\{\Delta_1(x,\theta) + \Delta_1'(x',\theta)\right\},
\end{aligned}$$

where we define

$$\Delta_1(x,\theta) = r(\theta_0)\|\nabla_\theta \mathcal{L}(\theta;x)\| + \frac{r(\theta_0)^2}{2}\sup_{\theta'\in\mathbb{B}(\theta_0,r(\theta_0))\cap\Theta}\left(\lambda_{\max}(H(\theta') - H(\theta';x))\right)_+,$$

and

$$\Delta_1'(x,\theta) = r(\theta_0)\|\nabla_\theta \mathcal{L}(\theta;x)\| + \frac{r(\theta_0)^2}{2}\sup_{\theta'\in\mathbb{B}(\theta_0,r(\theta_0))\cap\Theta}\left(\lambda_{\max}(H(\theta';x) - H(\theta'))\right)_+.$$

Applying this calculation with $x' = X$, we obtain

$$\frac{\frac{f(x;\theta)}{f(x;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot\,|\,\theta)}\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]} = \left(\mathbb{E}_{p_{\theta_0}(\cdot\,|\,\theta)}\left[\frac{\frac{f(X;\theta)}{f(X;\theta_0)}}{\frac{f(x;\theta)}{f(x;\theta_0)}}\right]\right)^{-1} \ge \frac{1}{\mathbb{E}_{p_{\theta_0}(\cdot\,|\,\theta)}\left[e^{\Delta_1'(X,\theta)}\right]\cdot e^{\Delta_1(x,\theta)}}$$

for all x and for all $\theta \in \Theta$ such that $\|\theta - \theta_0\| \le r(\theta_0)$. Returning to (A.3) and (A.4) above, and defining $\mathcal{E}_{\text{ball}}$ to be the event that $\|\widehat{\theta} - \theta_0\| \le r(\theta_0)$, we therefore have

$$\begin{aligned}
\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] &= \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[\left(1 - \frac{p_{\widehat{\theta}}(X \,|\, \widehat{\theta})}{p_{\theta_0}(X \,|\, \widehat{\theta})}\right)_+\right]\right] \\
&\le \mathbb{P}_{Q^*_{\theta_0}}\left(\mathcal{E}^c_{\text{ball}}\right) + \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[\mathbb{1}_{\mathcal{E}_{\text{ball}}} \cdot \left(1 - \frac{p_{\widehat{\theta}}(X \,|\, \widehat{\theta})}{p_{\theta_0}(X \,|\, \widehat{\theta})}\right)_+\right]\right] \\
&\le \mathbb{P}_{Q^*_{\theta_0}}\left(\mathcal{E}^c_{\text{ball}}\right) + \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[1 - \frac{1}{\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[e^{\Delta'_1(X,\widehat{\theta})}\right] \cdot e^{\Delta_1(X,\widehat{\theta})}}\right]\right] \\
&\le \mathbb{P}_{Q^*_{\theta_0}}\left(\mathcal{E}^c_{\text{ball}}\right) + \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[\Delta_1(X, \widehat{\theta})\right] + 1 - \frac{1}{\mathbb{E}_{p_{\theta_0}(\cdot \,|\, \widehat{\theta})}\left[e^{\Delta'_1(X,\widehat{\theta})}\right]}\right],
\end{aligned}$$

where the last step holds since $1 - ab \le (1-a) + (1-b) \le \log(1/a) + (1-b)$ for any $a, b \in (0, 1]$. (Note that, in the next-to-last line, the two random variables X appearing in the denominator are different—they are sampled independently conditional on $\widehat{\theta}$ from the distribution $p_{\theta_0}(\cdot \,|\, \widehat{\theta})$.)

Next, recall that by Lemma 1 together with the definition of $Q^*_{\theta_0}$, the joint distribution of $(X, \widehat{\theta})$ in this calculation above (i.e., $\widehat{\theta} \sim Q^*_{\theta_0}$ and $X \,|\, \widehat{\theta} \sim p_{\theta_0}(\cdot \,|\, \widehat{\theta})$), is equivalent to the joint distribution of $(X, \widehat{\theta}(X, W))$ when $(X, W) \sim P^*_{\theta_0}$. Therefore, our calculation above can be rewritten as follows (where we also apply Jensen's inequality to the last term):

$$\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] \le \mathbb{P}_{P^*_{\theta_0}}\left(\mathcal{E}^c_{\text{ball}}\right) + \mathbb{E}_{P^*_{\theta_0}}\left[\Delta_1(X, \widehat{\theta}(X, W))\right] + \left(1 - \frac{1}{\mathbb{E}_{P^*_{\theta_0}}\left[e^{\Delta'_1(X, \widehat{\theta}(X,W))}\right]}\right).$$

Next let

$$\Delta_2(x, w) = r(\theta_0)\sigma\|w\| + \frac{r(\theta_0)^2}{2} \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} \big(\lambda_{\max}(H(\theta) - H(\theta; x))\big)_+,$$

and

$$\Delta'_2(x, w) = r(\theta_0)\sigma\|w\| + \frac{r(\theta_0)^2}{2} \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} \big(\lambda_{\max}(H(\theta; x) - H(\theta))\big)_+,$$

and observe that $\Delta_1(x, \widehat{\theta}(x, w)) = \Delta_2(x, w)$ and $\Delta'_1(x, \widehat{\theta}(x, w)) = \Delta'_2(x, w)$ for all $(x, w) \in \Omega_{\text{SSOSP}}$, since $0 = \nabla_\theta \mathcal{L}(\widehat{\theta}(x, w); x, w) = \nabla_\theta \mathcal{L}(\widehat{\theta}(x, w); x) + \sigma w$ for all (x, w) in this set by definition. Therefore, since $(X, W) \in \Omega_{\text{SSOSP}}$ almost surely under $P^*_{\theta_0}$ by definition, we have

$$\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] \le \mathbb{P}_{P^*_{\theta_0}}\left(\mathcal{E}^c_{\text{ball}}\right) + \mathbb{E}_{P^*_{\theta_0}}\left[\Delta_2(X, W)\right] + \left(1 - \frac{1}{\mathbb{E}_{P^*_{\theta_0}}\left[e^{\Delta'_2(X,W)}\right]}\right).$$

Now let $\mathcal{E}_{\text{SSOSP}}$ be the event that $(X, W) \in \Omega_{\text{SSOSP}}$. Recall that $P^*_{\theta_0}$ is the joint distribution of $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditional on $\mathcal{E}_{\text{SSOSP}}$. Therefore, we can write this as follows where

we now take all probabilities and expectations with respect to $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$:

$$\begin{aligned}
&\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] \\
&\quad \le \mathbb{P}\left(\mathcal{E}^c_{\text{ball}} \mid \mathcal{E}_{\text{SSOSP}}\right) + \mathbb{E}\left[\Delta_2(X, W) \mid \mathcal{E}_{\text{SSOSP}}\right] + \left(1 - \frac{1}{\mathbb{E}\left[e^{\Delta_2'(X,W)} \mid \mathcal{E}_{\text{SSOSP}}\right]}\right) \\
&\quad \le \frac{\mathbb{P}\left(\mathcal{E}^c_{\text{ball}} \cap \mathcal{E}_{\text{SSOSP}}\right) + \mathbb{E}\left[\Delta_2(X, W)\right]}{\mathbb{P}\left(\mathcal{E}_{\text{SSOSP}}\right)} + \left(1 - \frac{\mathbb{P}\left(\mathcal{E}_{\text{SSOSP}}\right)}{\mathbb{E}\left[e^{\Delta_2'(X,W)} \cdot \mathbb{1}_{\mathcal{E}_{\text{SSOSP}}}\right]}\right) \\
&\quad \le \frac{\mathbb{P}\left(\mathcal{E}^c_{\text{ball}} \cap \mathcal{E}_{\text{SSOSP}}\right) + \mathbb{E}\left[\Delta_2(X, W)\right]}{\mathbb{P}\left(\mathcal{E}_{\text{SSOSP}}\right)} + \left(1 - \frac{1 - \mathbb{P}\left(\mathcal{E}^c_{\text{SSOSP}}\right)}{\mathbb{E}\left[e^{\Delta_2'(X,W)}\right] - \mathbb{P}\left(\mathcal{E}^c_{\text{SSOSP}}\right)}\right) \\
&\quad \le \frac{\mathbb{P}\left(\mathcal{E}^c_{\text{ball}} \cap \mathcal{E}_{\text{SSOSP}}\right) + \mathbb{E}\left[\Delta_2(X, W)\right] + \log \mathbb{E}\left[e^{\Delta_2'(X,W)}\right]}{\mathbb{P}\left(\mathcal{E}_{\text{SSOSP}}\right)},
\end{aligned}$$

where the next-to-last step holds since $\Delta_2'(X, W) \ge 0$ by definition, and the last step holds since $1 - \frac{1-a}{b-a} \le \frac{1-1/b}{1-a} \le \frac{\log(b)}{1-a}$ for all $a \in [0, 1)$ and $b \ge 1$. Finally, we apply our assumptions. By Assumption 2, we have $\mathbb{P}\left(\mathcal{E}_{\text{ball}} \cap \mathcal{E}_{\text{SSOSP}}\right) \ge 1 - \delta(\theta_0)$, and so

$$\mathbb{P}\left(\mathcal{E}^c_{\text{ball}} \cap \mathcal{E}_{\text{SSOSP}}\right) \le \delta(\theta_0) - \mathbb{P}\left(\mathcal{E}^c_{\text{SSOSP}}\right).$$

Next,

$$\begin{aligned}
\mathbb{E}\left[\Delta_2(X, W)\right] = \mathbb{E}\left[r(\theta_0)\sigma\|W\|\right] + \mathbb{E}\left[\frac{r(\theta_0)^2}{2} \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} \big(\lambda_{\max}(H(\theta) - H(\theta; X))\big)_+\right] \\
\le \frac{1}{2}\log \mathbb{E}\left[e^{2r(\theta_0)\sigma\|W\|}\right] + \frac{\varepsilon(\theta_0)}{2},
\end{aligned}$$

where the last step holds by Jensen's inequality for the first term and by the bound (3.2) in Assumption 3 for the second term. And, by Cauchy–Schwarz,

$$\begin{aligned}
\log \mathbb{E}\left[e^{\Delta_2'(X,W)}\right] \le \frac{1}{2}\log \mathbb{E}\left[e^{2r(\theta_0)\sigma\|W\|}\right] + \frac{1}{2}\log \mathbb{E}\left[e^{r(\theta_0)^2 \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} \left(\lambda_{\max}(H(\theta;X) - H(\theta))\right)_+}\right] \\
\le \frac{1}{2}\log \mathbb{E}\left[e^{2r(\theta_0)\sigma\|W\|}\right] + \frac{\varepsilon(\theta_0)}{2},
\end{aligned}$$

where the last step holds by the bound (3.3) in Assumption 3. Finally, since $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ we know that $\mathbb{E}\left[e^{t\|W\|}\right] \le e^{t + t^2/2d}$ for any $t > 0$ (see, e.g., [Boucheron et al., 2013, Theorem 5.5]). Therefore,

$$\log \mathbb{E}\left[e^{2r(\theta_0)\sigma\|W\|}\right] \le 2\sigma \cdot r(\theta_0) + \frac{2\sigma^2 \cdot r(\theta_0)^2}{d} \le 3\sigma \cdot r(\theta_0),$$

where the last step holds since $d \ge 1$ and we can assume $2\sigma \cdot r(\theta_0) \le 1$ (as otherwise, the result of the theorem holds trivially). Combining everything, we have

$$\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot \,|\, \widehat{\theta}), p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})\big)\right] \le \frac{3\sigma \cdot r(\theta_0) + \delta(\theta_0) + \varepsilon(\theta_0) - \mathbb{P}\left(\mathcal{E}^c_{\text{SSOSP}}\right)}{1 - \mathbb{P}\left(\mathcal{E}^c_{\text{SSOSP}}\right)}.$$

Since total variation distance is bounded by 1, trivially we can relax this to

$$\mathbb{E}_{Q^*_{\theta_0}}\left[\mathsf{d}_{\mathsf{TV}}\big(p_{\theta_0}(\cdot\,|\,\widehat{\theta}), p_{\widehat{\theta}}(\cdot\,|\,\widehat{\theta})\big)\right] \leq 3\sigma \cdot r(\theta_0) + \delta(\theta_0) + \varepsilon(\theta_0).$$

Returning to (A.2), we see that the aCSS procedure (2.8) satisfies

$$\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq 3\sigma \cdot r(\theta_0) + \delta(\theta_0) + \varepsilon(\theta_0),$$

as desired.

A.2 Proof of Lemma 1

Consider the joint distribution $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditioned on the event that $(X, W) \in \Omega_{\mathrm{SSOSP}}$, which is assumed to occur with positive probability. The joint density of (X, W), after conditioning on this event, is therefore proportional to the function

$$g_{\theta_0}(x, w) = f(x; \theta_0) \cdot \exp\left\{-\tfrac{d}{2}\|w\|^2\right\} \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}, \tag{A.5}$$

with respect to the measure $\nu_{\mathcal{X}} \times \mathrm{Leb}$. We will consider the induced joint distribution of $(X, \widehat{\theta}(X, W))$, and will calculate its joint density.

Define ψ and ψ^{-1} as in Lemma 2. Fix any measurable subset $A \subseteq \Psi_{\mathrm{SSOSP}}$. Then, writing $\psi^{-1}(A) = \{(x, w) \in \Omega_{\mathrm{SSOSP}} : \psi(x, w) \in A\} \subseteq \Omega_{\mathrm{SSOSP}}$,

$$\mathbb{P}\left((X, \widehat{\theta}(X, W)) \in A\right) = \mathbb{P}\left((X, W) \in \psi^{-1}(A)\right) = \frac{\int_{\psi^{-1}(A)} g_{\theta_0}(x, w)\ \mathsf{d}\nu_{\mathcal{X}}(x)\mathsf{d}w}{\int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x', w')\ \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'},$$

where the probability is taken with respect to $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditioned on the event that $(X, W) \in \Omega_{\mathrm{SSOSP}}$. (Note that, since g_{θ_0} is proportional to a density on (X, W) with respect to $\nu_{\mathcal{X}} \times \mathrm{Leb}$, this implies that the denominator $\int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x', w')\ \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'$ in the last expression above must be finite and positive.)

From this point on, the result essentially follows from a change-of-variables calculation, under the transformation $\theta = \widehat{\theta}(x, w)$. However, with our weak assumptions, we cannot assume standard conditions (such as, e.g., the support of $\widehat{\theta}\,|\,X$ being an open subset of $\mathbb{R}^d$—it may even be the case that this set does not contain any open subset), so we will need to be careful. Fixing any $x \in \mathcal{X}$, a change-of-variables calculation, proved formally in Appendix B.4 below, establishes that

$$\begin{aligned}
&\int_\Theta \exp\left\{-\tfrac{d}{2\sigma^2}\|\nabla_\theta \mathcal{L}(\theta; x)\|^2\right\} \cdot \det(\nabla^2_\theta \mathcal{L}(\theta; x)) \cdot \mathbb{1}_{(x,\theta)\in A\cap\Psi_{\mathrm{SSOSP}}}\ \mathsf{d}\theta \\
&\quad = \sigma^d \int_{\mathbb{R}^d} \exp\left\{-\tfrac{d}{2\sigma^2}\|\nabla_\theta \mathcal{L}(\widehat{\theta}(x, w); x)\|^2\right\} \cdot \mathbb{1}_{(x,\widehat{\theta}(x,w))\in A} \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\ \mathsf{d}w \qquad (A.6) \\
&\quad = \sigma^d \int_{\mathbb{R}^d} \exp\left\{-\tfrac{d}{2}\|w\|^2\right\} \cdot \mathbb{1}_{(x,\widehat{\theta}(x,w))\in A} \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\ \mathsf{d}w \\
&\quad = \sigma^d \int_{\mathbb{R}^d} \exp\left\{-\tfrac{d}{2}\|w\|^2\right\} \cdot \mathbb{1}_{(x,w)\in\psi^{-1}(A)\cap\Omega_{\mathrm{SSOSP}}}\ \mathsf{d}w, \qquad (A.7)
\end{aligned}$$

where the second step uses the fact that $w = -\frac{\nabla_\theta \mathcal{L}(\widehat{\theta}(x,w);x)}{\sigma}$ for any $(x,w) \in \Omega_{\text{SSOSP}}$ by the SSOSP conditions, and the last step applies the definition of ψ as in Lemma 2. Now define the function

$$h_{\theta_0}(x,\theta) := \frac{f(x;\theta_0)\exp\left\{-\frac{d}{2\sigma^2}\|\nabla_\theta \mathcal{L}(\theta;x)\|^2\right\} \cdot \det(\nabla^2_\theta \mathcal{L}(\theta;x)) \cdot \mathbb{1}_{x\in\mathcal{X}_\theta}}{\sigma^d \int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'}$$

on $(x,\theta) \in \mathcal{X}\times\Theta$. We then have

$$\begin{aligned}
&\mathbb{P}\left((X,\widehat{\theta}(X,W)) \in A\right) = \frac{\int_{\psi^{-1}(A)} g_{\theta_0}(x,w)\, \mathsf{d}\nu_{\mathcal{X}}(x)\mathsf{d}w}{\int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'} \\
&= \frac{\int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x;\theta_0)\cdot\exp\left\{-\frac{d}{2}\|w\|^2\right\}\cdot \mathbb{1}_{(x,w)\in\psi^{-1}(A)\cap\Omega_{\text{SSOSP}}}\, \mathsf{d}w\, \mathsf{d}\nu_{\mathcal{X}}(x)}{\int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'} \quad \text{by (A.5)} \\
&= \frac{\int_{\mathcal{X}} f(x;\theta_0)\int_\Theta \exp\left\{-\frac{d}{2\sigma^2}\|\nabla_\theta \mathcal{L}(\theta;x)\|^2\right\}\cdot\det(\nabla^2_\theta\mathcal{L}(\theta;x))\cdot\mathbb{1}_{(x,\theta)\in A\cap\Psi_{\text{SSOSP}}}\, \mathsf{d}\theta\, \mathsf{d}\nu_{\mathcal{X}}(x)}{\sigma^d\int_{\mathcal{X}\times\mathbb{R}^d} g_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\mathsf{d}w'} \quad \text{by (A.7)} \\
&= \int_A h_{\theta_0}(x,\theta)\, \mathsf{d}\nu_{\mathcal{X}}(x)\, \mathsf{d}\theta,
\end{aligned}$$

where the last step holds since $\mathbb{1}_{x\in\mathcal{X}_\theta} = \mathbb{1}_{(x,\theta)\in\Psi_{\text{SSOSP}}}$ for all (x,θ), by definition of Ψ_{SSOSP}. Therefore, this calculation establishes that, conditional on the event that $\widehat{\theta}(X,W)$ is a SSOSP of $\mathcal{L}(\theta;X,W)$, the joint distribution of $(X,\widehat{\theta}(X,W))$ has density $h_{\theta_0}(x,\theta)$ with respect to the base measure $\nu_{\mathcal{X}} \times \text{Leb}$.

Finally, since $h_{\theta_0}(x,\theta)$ is the joint density of $(X,\widehat{\theta}) = (X,\widehat{\theta}(X,W))$, we therefore see that $X \,|\, \widehat{\theta}$ has conditional density equal to

$$\frac{h_{\theta_0}(x,\widehat{\theta})}{\int_{x'} h_{\theta_0}(x',\widehat{\theta})\, \mathsf{d}\nu_{\mathcal{X}}(x')} \propto f(x;\theta_0)\exp\left\{-\frac{d}{2\sigma^2}\|\nabla_\theta\mathcal{L}(\widehat{\theta};x)\|^2\right\}\cdot\det(\nabla^2_\theta\mathcal{L}(\widehat{\theta};x))\cdot\mathbb{1}_{x\in\mathcal{X}_{\widehat{\theta}}},$$

which verifies the desired expression (2.4).

B Additional proofs

B.1 Proof of Lemma 2

First we check that ψ is injective on Ω_{SSOSP}, which holds since for any (x,θ), if $\psi(x',w) = (x,\theta)$ then we must have $x = x'$ trivially and we must have $w = -\frac{\nabla_\theta\mathcal{L}(\theta;x)}{\sigma}$ by definition of the SSOSP conditions. This establishes that ψ is injective and that the inverse function (on the image of ψ) is given by $\psi^{-1}(x,\theta) = \left(x, -\frac{\nabla_\theta\mathcal{L}(\theta;x)}{\sigma}\right)$ as claimed above.

Now we verify that Ψ_{SSOSP} is the image of ψ. Fix any $(x,\theta) \in \mathcal{X}\times\Theta$. First, suppose $(x,\theta) \in \psi(\Omega_{\text{SSOSP}})$, i.e., we have $\theta = \widehat{\theta}(x,w)$ for some w such that $(x,w) \in \Omega_{\text{SSOSP}}$. Then by definition of Ω_{SSOSP}, θ is a SSOSP of $\mathcal{L}(\theta;x,w)$, and so $x \in \mathcal{X}_\theta$ and therefore $(x,\theta) \in \Psi_{\text{SSOSP}}$. Conversely suppose that $(x,\theta) \in \Psi_{\text{SSOSP}}$. Then by definition, $x \in \mathcal{X}_\theta$ and so there exists some w such that $\theta = \widehat{\theta}(x,w)$ and θ is a SSOSP of $\mathcal{L}(\theta;x,w)$. Therefore, for this choice of w, we have $(x,w) \in \Omega_{\text{SSOSP}}$ and so $(x,\theta) = \psi(x,w) \in \psi(\Omega_{\text{SSOSP}})$.

B.2 Distance to exchangeability for mixture distributions

In this section, we verify the claim (A.1) that appears in the proof of Theorem 1. Specifically, we need to show that the distance-to-exchangeability $\mathsf{d}_{\mathsf{exch}}$ introduced in Definition 1 is convex on the space of distributions.

Lemma 3. *Consider any distributions P_0, P_1 on $(A_1, \ldots, A_k)$, and any $c \in [0,1]$. Let $P = (1-c) \cdot P_0 + c \cdot P_1$ be the mixture distribution. Then*

$$\mathsf{d}_{\mathsf{exch}}(P) \leq (1-c) \cdot \mathsf{d}_{\mathsf{exch}}(P_0) + c \cdot \mathsf{d}_{\mathsf{exch}}(P_1).$$

With this lemma in place, we have

$$\begin{aligned}\mathsf{d}_{\mathsf{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \leq \mathbb{P}\left((X,W) \in \Omega_{\text{SSOSP}}\right) \cdot \mathsf{d}_{\mathsf{exch}}\left(\begin{array}{c}\text{Distrib. of } X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \\ \text{condl. on } (X,W) \in \Omega_{\text{SSOSP}}\end{array}\right) \\ + \mathbb{P}\left((X,W) \notin \Omega_{\text{SSOSP}}\right) \cdot \mathsf{d}_{\mathsf{exch}}\left(\begin{array}{c}\text{Distrib. of } X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \\ \text{condl. on } (X,W) \notin \Omega_{\text{SSOSP}}\end{array}\right).\end{aligned}$$

Furthermore, we know that

$$\mathsf{d}_{\mathsf{exch}}\left(\begin{array}{c}\text{Distrib. of } X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \\ \text{condl. on } (X,W) \notin \Omega_{\text{SSOSP}}\end{array}\right) = 0$$

since, on the event that $(X,W) \notin \Omega_{\text{SSOSP}}$, we set $\widetilde{X}^{(1)} = \cdots = \widetilde{X}^{(M)} = X$ by definition of the method. Therefore, the claim (A.1) must hold.

Proof of Lemma 3. Fix any $\varepsilon > 0$. By definition of $\mathsf{d}_{\mathsf{exch}}$, for each $\ell = 0, 1$ we can find some exchangeable distribution Q_ℓ on $(B_1, \ldots, B_k)$ such that

$$\mathsf{d}_{\mathsf{TV}}(P_\ell, Q_\ell) \leq \mathsf{d}_{\mathsf{exch}}(P_\ell) + \varepsilon.$$

Next define the mixture distribution $Q = (1-c) \cdot Q_0 + c \cdot Q_1$. Clearly Q is exchangeable, inheriting this property from Q_0 and Q_1, and therefore $\mathsf{d}_{\mathsf{exch}}(P) \leq \mathsf{d}_{\mathsf{TV}}(P, Q)$. Furthermore, for any measurable subset A, we have

$$\begin{aligned}&\left|P(A) - Q(A)\right| = \left|\left((1-c) \cdot P_0(A) + c \cdot P_1(A)\right) - \left((1-c) \cdot Q_0(A) + c \cdot Q_1(A)\right)\right| \\ &\leq (1-c) \cdot |P_0(A) - Q_0(A)| + c \cdot |P_1(A) - Q_1(A)| \leq (1-c) \cdot \mathsf{d}_{\mathsf{TV}}(P_0, Q_0) + c \cdot \mathsf{d}_{\mathsf{TV}}(P_1, Q_1).\end{aligned}$$

This establishes that $\mathsf{d}_{\mathsf{TV}}(P, Q) \leq (1-c) \cdot \mathsf{d}_{\mathsf{TV}}(P_0, Q_0) + c \cdot \mathsf{d}_{\mathsf{TV}}(P_1, Q_1)$, and therefore,

$$\mathsf{d}_{\mathsf{exch}}(P) \leq (1-c) \cdot (\mathsf{d}_{\mathsf{exch}}(P_0) + \varepsilon) + c \cdot (\mathsf{d}_{\mathsf{exch}}(P_1) + \varepsilon).$$

Since $\varepsilon > 0$ can be taken to be arbitrarily small, this proves the lemma. □

B.3 Verifying that (2.7) defines a density

To ensure that our procedure is well defined, we need to check that

$$p_{\widehat{\theta}}(x \,|\, \widehat{\theta}) \propto p^{\mathrm{un}}_{\widehat{\theta}}(x)$$

defines a valid density with respect to $\nu_{\mathcal{X}}$, where the unnormalized function is given by

$$p^{\mathrm{un}}_{\theta}(x) := f(x;\theta) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\theta;x)\|^2}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\theta;x)\right) \cdot \mathbb{1}_{x\in\mathcal{X}_\theta}.$$

The following lemma verifies all the necessary conditions:

Lemma 4. *If Assumptions 1 and 3 hold, then for all $\theta \in \Theta$ the function $x \mapsto p^{\mathrm{un}}_{\theta}(x)$ is nonnegative and integrable with respect to $\nu_{\mathcal{X}}$. Furthermore, if the event that $\widehat{\theta} = \widehat{\theta}(X,W)$ is a SSOSP of $\mathcal{L}(\theta;X,W)$ has positive probability, then conditional on this event,*

$$\int_{\mathcal{X}} p^{\mathrm{un}}_{\widehat{\theta}}(x)\, \mathsf{d}\nu_{\mathcal{X}}(x) > 0.$$

holds almost surely.

Proof. First we check nonnegativity. For any θ and any x, $f(x;\theta) > 0$ by Assumption 1. Furthermore, if $x \in \mathcal{X}_\theta$ then $\nabla^2_\theta \mathcal{L}(\theta;x) \succ 0$ and so $\det(\nabla^2_\theta \mathcal{L}(\theta;x)) > 0$ by definition of the SSOSP conditions. This verifies that $p^{\mathrm{un}}_{\theta}(x) \geq 0$ for all (x,θ). Next we check integrability. We have

$$\begin{aligned}
&\int_{\mathcal{X}} p^{\mathrm{un}}_{\widehat{\theta}}(x)\, \mathsf{d}\nu_{\mathcal{X}}(x) \leq \int_{\mathcal{X}} f(x;\theta) \cdot \det\left(\nabla^2_\theta \mathcal{L}(\theta;x)\right) \cdot \mathbb{1}_{\nabla^2_\theta \mathcal{L}(\theta;x)\succ 0}\, \mathsf{d}\nu_{\mathcal{X}}(x) \\
&\leq \int_{\mathcal{X}} f(x;\theta) \cdot \left(\lambda_{\max}(\nabla^2_\theta \mathcal{L}(\theta;x))\right)^d_+ \, \mathsf{d}\nu_{\mathcal{X}}(x) \\
&\leq \frac{d!}{r(\theta)^{2d}} \int_{\mathcal{X}} f(x;\theta) \cdot \exp\left\{r(\theta)^2 \left(\lambda_{\max}(\nabla^2_\theta \mathcal{L}(\theta;x))\right)_+\right\}\, \mathsf{d}\nu_{\mathcal{X}}(x) \\
&\leq \frac{d!}{r(\theta)^{2d}} \int_{\mathcal{X}} f(x;\theta) \cdot \exp\left\{r(\theta)^2 \left(\lambda_{\max}(H(\theta;x) - H(\theta)\right)_+ + r(\theta)^2 \left(\lambda_{\max}(H(\theta) + \nabla^2_\theta \mathcal{R}(\theta)\right)_+\right\}\, \mathsf{d}\nu_{\mathcal{X}}(x) \\
&\leq \frac{d!}{r(\theta)^{2d}} \cdot e^{\varepsilon(\theta)} \cdot \exp\left\{r(\theta)^2 \left(\lambda_{\max}(H(\theta) + \nabla^2_\theta \mathcal{R}(\theta)\right)_+\right\},
\end{aligned}$$

where the last step holds by Assumption 3. This proves that $\int_{\mathcal{X}} p^{\mathrm{un}}_{\widehat{\theta}}(x)\, \mathsf{d}\nu_{\mathcal{X}}(x)$ is finite.

Finally we check that $\int_{\mathcal{X}} p^{\mathrm{un}}_{\widehat{\theta}}(x)\, \mathsf{d}\nu_{\mathcal{X}}(x) > 0$ almost surely. Since $f(x;\theta) > 0$ for all x, θ by Assumption 1, it is equivalent to verify that $\int_{\mathcal{X}} \frac{f(x;\theta_0)}{f(x;\widehat{\theta})} p^{\mathrm{un}}_{\widehat{\theta}}(x)\, \mathsf{d}\nu_{\mathcal{X}}(x) > 0$ almost surely. Recalling from (2.4) that $p_{\theta_0}(x|\widehat{\theta}) \propto \frac{f(x;\theta_0)}{f(x;\widehat{\theta})} \cdot p^{\mathrm{un}}_{\widehat{\theta}}(x)$ is the conditional density of $X \,|\, \widehat{\theta}$, this must be true. □

B.4 Change of variables calculation

In this section, we verify the change-of-variables calculation needed in the proof of Lemma 1. Specifically, the step (A.6) follows by applying the lemma below to the function

$$\rho(x,\theta) = \exp\left\{-\tfrac{d}{2\sigma^2}\|\nabla_\theta \mathcal{L}(\widehat{\theta}(x,w);x)\|^2\right\} \cdot \mathbb{1}_{(x,\theta)\in A}.$$

Lemma 5. *Suppose Assumption 1 holds. For all nonnegative measurable functions $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$, it holds for all $x \in \mathcal{X}$ that*

$$\int_\Theta \rho(x,\theta) \cdot \det(\nabla^2_\theta \mathcal{L}(\theta;x)) \cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\, \mathsf{d}\theta = \sigma^d \int_{\mathbb{R}^d} \rho(x,\widehat{\theta}(x,w)) \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\, \mathsf{d}w.$$

Proof of Lemma 5. Define

$$A_x = \{\theta \in \Theta : \nabla^2 \mathcal{L}(\theta;x) \succ 0\},$$

which is an open set since $\mathcal{L}(\theta;x)$ is continuously twice differentiable in θ. If this set is empty then the lemma is trivial (since the left- and right-hand side are both equal to zero), so from this point on we will assume A_x is nonempty. Let

$$B_x = \{w \in \mathbb{R}^d : \widehat{\theta}(x,w) \in A_x\}.$$

By definition, if $(x,\theta) \in \Psi_{\mathrm{SSOSP}}$ then we must have $\theta \in A_x$, and similarly if $(x,w) \in \Omega_{\mathrm{SSOSP}}$ then we must have $\widehat{\theta}(x,w) \in A_x$ and so $w \in B_x$. Therefore, to prove the lemma, it is sufficient to show that

$$\int_{A_x} \rho(x,\theta) \cdot \det(\nabla^2_\theta \mathcal{L}(\theta;x)) \cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\, \mathsf{d}\theta = \sigma^d \int_{B_x} \rho(x,\widehat{\theta}(x,w)) \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\, \mathsf{d}w. \quad \text{(B.1)}$$

Next define nested sets

$$A_{x,\lambda} = \{\theta \in A_x : \mathbb{B}(\theta,\lambda) \subseteq \Theta \text{ and } \nabla^2 \mathcal{L}(\theta';x) \succ 0 \text{ for all } \theta' \in \mathbb{B}(\theta,\lambda)\}$$

indexed by $\lambda > 0$. Since Θ is an open subset of $\mathbb{R}^d$, and $\mathcal{L}(\theta;x)$ is continuously twice differentiable in θ, we see that $A_x = \cup_{\lambda>0} A_{x,\lambda}$. Similarly we have $B_x = \cup_{\lambda>0} B_{x,\lambda}$ where

$$B_{x,\lambda} = \{w \in \mathbb{R}^d : \widehat{\theta}(x,w) \in A_{x,\lambda}\}.$$

By the monotone convergence theorem, this implies that

$$\int_{A_x} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot\mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\, \mathsf{d}\theta = \lim_{\lambda\to 0} \int_{A_{x,\lambda}} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot\mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\, \mathsf{d}\theta,$$

and similarly,

$$\int_{B_x} \rho(x,\widehat{\theta}(x,w)) \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\, \mathsf{d}w = \lim_{\lambda\to 0} \int_{B_{x,\lambda}} \rho(x,\widehat{\theta}(x,w)) \cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\, \mathsf{d}w.$$

Therefore, to prove (B.1), it is sufficient to show that, for each $\lambda > 0$,

$$\int_{A_{x,\lambda}} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\,\mathsf{d}\theta = \sigma^d \int_{B_{x,\lambda}} \rho(x,\widehat{\theta}(x,w))\cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\,\mathsf{d}w. \quad \text{(B.2)}$$

From this point on we will treat $\lambda > 0$ as fixed. Let $S_1, S_2, \ldots$ be a countable collection of disjoint open sets, each of diameter $\le \lambda$, such that $\mathrm{Leb}\big(\mathbb{R}^d \backslash (\cup_{k\ge 1} S_k)\big) = 0$ (for example, we can partition $\mathbb{R}^d$ into countably many sufficiently small hypercubes). Then

$$\int_{A_{x,\lambda}} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\,\mathsf{d}\theta = \sum_{k\ge 1}\int_{A_{x,\lambda,k}} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\,\mathsf{d}\theta,$$

where $A_{x,\lambda,k} = A_{x,\lambda} \cap S_k$, and similarly

$$\int_{B_{x,\lambda}} \rho(x,\widehat{\theta}(x,w))\cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\,\mathsf{d}w = \sum_{k\ge 1}\int_{B_{x,\lambda,k}} \rho(x,\widehat{\theta}(x,w))\cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\,\mathsf{d}w,$$

where

$$B_{x,\lambda,k} = \{w \in \mathbb{R}^d : \widehat{\theta}(x,w) \in A_{x,\lambda,k}\}.$$

Therefore, to prove (B.2), it is sufficient to show that, for each $\lambda > 0$ and each $k \ge 1$,

$$\int_{A_{x,\lambda,k}} \rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x))\cdot \mathbb{1}_{(x,\theta)\in\Psi_{\mathrm{SSOSP}}}\,\mathsf{d}\theta = \sigma^d \int_{B_{x,\lambda,k}} \rho(x,\widehat{\theta}(x,w))\cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\,\mathsf{d}w. \quad \text{(B.3)}$$

From this point on we will treat both $\lambda > 0$ and $k \ge 1$ as fixed, and will prove (B.3). First, by definition of the SSOSP conditions, if $(x,\theta) \in \Psi_{\mathrm{SSOSP}}$ then we must have

$$\theta = \widehat{\theta}(x,\phi_x(\theta)) \quad \text{where} \quad \phi_x(\theta) := -\frac{\nabla_\theta \mathcal{L}(\theta;x)}{\sigma},$$

and furthermore, $\nabla^2_\theta \mathcal{L}(\theta;x) \succ 0$ and so $\det(\nabla^2_\theta \mathcal{L}(\theta;x)) > 0$. Therefore, we can calculate that

$$\rho(x,\theta)\cdot\det(\nabla^2_\theta \mathcal{L}(\theta;x)) = \rho(x,\widehat{\theta}(x,\phi_x(\theta)))\cdot\sigma^d|\det(\nabla_\theta\phi_x(\theta))|$$

for all $(x,\theta) \in \Psi_{\mathrm{SSOSP}}$, and so (B.3) is equivalent to the claim that

$$\int_{A_{x,\lambda,k}} \rho(x,\widehat{\theta}(x,\phi_x(\theta)))\cdot|\det(\nabla_\theta\phi_x(\theta))|\cdot \mathbb{1}_{(x,\widehat{\theta}(x,\phi_x(\theta)))\in\Psi_{\mathrm{SSOSP}}}\,\mathsf{d}\theta = \int_{B_{x,\lambda,k}} \rho(x,\widehat{\theta}(x,w))\cdot \mathbb{1}_{(x,w)\in\Omega_{\mathrm{SSOSP}}}\,\mathsf{d}w. \quad \text{(B.4)}$$

Next, we show that $\phi_x : A_{x,\lambda,k} \to \phi_x(A_{x,\lambda,k})$ is a diffeomorphism. ϕ_x is clearly differentiable, and its derivative is invertible since $\nabla_\theta\phi_x(\theta) = (-\sigma)^{-d}\nabla^2_\theta \mathcal{L}(\theta;x)$, and $\nabla^2_\theta \mathcal{L}(\theta;x) \succ 0$ on $A_{x,\lambda,k}$ by definition. To check injectivity, if $\phi_x(\theta) = \phi_x(\theta')$ for some $\theta, \theta' \in A_{x,\lambda,k}$, then by

Taylor's theorem we must have $\nabla^2_\theta \mathcal{L}((1-t)\theta + t\theta'; x) \cdot (\theta' - \theta) = 0$ for some $t \in [0,1]$. Since the diameter of S_k (and therefore, of $A_{x,\lambda,k}$) is $\leq \lambda$, we must have $\|\theta - \theta'\| \leq \lambda$ and therefore $(1-t)\theta + t\theta' \in \mathbb{B}(\theta, \lambda)$. By definition of $A_{x,\lambda}$, this implies that $\nabla^2_\theta \mathcal{L}((1-t)\theta + t\theta'; x) \succ 0$, and we conclude that $\theta' - \theta = 0$, thus establishing injectivity. Therefore, $\phi_x : A_{x,\lambda,k} \to \phi_x(A_{x,\lambda,k})$ is a diffeomorphism. Since $A_{x,\lambda,k} \subseteq \mathbb{R}^d$ is an open set, by the change-of-variables formula we therefore have

$$\int_{A_{x,\lambda,k}} \rho(x, \widehat{\theta}(x, \phi_x(\theta))) \cdot |\det(\nabla_\theta \phi_x(\theta))| \cdot \mathbb{1}_{(x, \widehat{\theta}(x, \phi_x(\theta))) \in \Psi_{\mathrm{SSOSP}}} \, \mathsf{d}\theta$$
$$= \int_{\phi_x(A_{x,\lambda,k})} \rho(x, \widehat{\theta}(x, w)) \cdot \mathbb{1}_{(x, \widehat{\theta}(x,w)) \in \Psi_{\mathrm{SSOSP}}} \, \mathsf{d}w,$$

Therefore, to prove (B.4), we now only need to check that

$$\mathbb{1}\left\{w \in \phi_x(A_{x,\lambda,k}),\ (x, \widehat{\theta}(x,w)) \in \Psi_{\mathrm{SSOSP}}\right\} = \mathbb{1}\left\{w \in B_{x,\lambda,k},\ (x,w) \in \Omega_{\mathrm{SSOSP}}\right\}$$

for all (x, w). First suppose $w \in \phi_x(A_{x,\lambda,k})$ and $(x, \widehat{\theta}(x,w)) \in \Psi_{\mathrm{SSOSP}}$. Then we have $w = \phi_x(\theta)$ for some $\theta \in A_{x,\lambda,k}$. By definition, this means $(x, \theta) \in \Psi_{\mathrm{SSOSP}}$, and so we must have some w' such that $\theta = \widehat{\theta}(x, w')$ and θ is a SSOSP of $\mathcal{L}(\theta; x, w')$. By the SSOSP conditions, this implies that $0 = \nabla_\theta \mathcal{L}(\theta; x, w')$ and so $w' = \phi_x(\theta)$, and therefore $w = w'$. Therefore, $(x,w) \in \Omega_{\mathrm{SSOSP}}$, and $\widehat{\theta}(x,w) \in A_{x,\lambda,k}$ which implies $w \in B_{x,\lambda,k}$. Conversely, suppose that $w \in B_{x,\lambda,k}$ and $(x,w) \in \Omega_{\mathrm{SSOSP}}$. Then by definition of $B_{x,\lambda,k}$, we have $\widehat{\theta}(x,w) \in A_{x,\lambda,k}$. Furthermore, by the SSOSP conditions we must have $0 = \nabla_\theta \mathcal{L}(\widehat{\theta}(x,w); x, w)$ and so $w = \phi_x(\widehat{\theta}(x,w))$, and therefore, $w \in \phi_x(A_{x,\lambda,k})$ and $(x, \widehat{\theta}(x,w)) \in \Psi_{\mathrm{SSOSP}}$. This completes the proof of (B.4), and therefore proves the lemma. □

C Proofs for examples

We now turn to establishing that our examples all satisfy the assumptions needed for aCSS to control Type I error. The regularity conditions (Assumption 1) hold by definition for all of our examples, so we only need to verify the properties of the estimator $\widehat{\theta}$ (Assumption 2) and the Hessian conditions (Assumption 3).

C.1 Checking Assumption 3

The Hessian conditions (3.2) and (3.3) are immediately implied by the stronger condition

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot \|H(\theta; X) - H(\theta)\|\right\}\right] \leq e^{\varepsilon(\theta_0)}. \tag{C.1}$$

We will check that this stronger condition holds for each of our examples. Specifically, fixing $\theta_0 \in \Theta$ we will prove that, for any $c > 0$ we can find $c' > 0$ such that

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta \in \mathbb{B}(\theta_0, c\sqrt{\frac{\log n}{n}}) \cap \Theta} \frac{c^2 \log n}{n} \cdot \|H(\theta; X) - H(\theta)\|\right\}\right] \leq e^{c'\varepsilon_n} \tag{C.2}$$

for all sufficiently large n, where ε_n is some vanishing term (specified below for each example) that does not depend on our choice of c. Since later on we will verify that Assumption 2 holds with $r(\theta_0) \asymp \sqrt{\frac{\log n}{n}}$, this will be sufficient to verify that (C.1) holds.

C.1.1 Checking Assumption 3 for Example 1

For the canonical GLM setting (Example 1), we can calculate

$$H(\theta; x) = \sum_{i=1}^{n} Z_i Z_i^\top \cdot a''(Z_i^\top \theta),$$

which does not depend on x. Therefore, $H(\theta) = H(\theta; x)$ for all x, or in other words, $\|H(\theta; X) - H(\theta)\| = 0$ almost surely. Therefore (C.2) holds trivially with $\varepsilon_n = 0$.

C.1.2 Checking Assumption 3 for Example 2

For the Behrens–Fisher problem (Example 2), we can calculate

$$H(\theta; x) = \begin{pmatrix} \frac{n^{(0)}}{\gamma^{(0)}} + \frac{n^{(1)}}{\gamma^{(1)}} & \frac{\sum_{i=1}^{n^{(0)}} (x_i^{(0)} - \mu)}{(\gamma^{(0)})^2} & \frac{\sum_{i=1}^{n^{(1)}} (x_i^{(1)} - \mu)}{(\gamma^{(1)})^2} \\ \frac{\sum_{i=1}^{n^{(0)}} (x_i^{(0)} - \mu)}{(\gamma^{(0)})^2} & -\frac{n^{(0)}}{2(\gamma^{(0)})^2} + \sum_{i=1}^{n^{(0)}} \frac{(x_i^{(0)} - \mu)^2}{(\gamma^{(0)})^3} & 0 \\ \frac{\sum_{i=1}^{n^{(1)}} (x_i^{(1)} - \mu)}{(\gamma^{(1)})^2} & 0 & -\frac{n^{(1)}}{2(\gamma^{(1)})^2} + \sum_{i=1}^{n^{(1)}} \frac{(x_i^{(1)} - \mu)^2}{(\gamma^{(1)})^3} \end{pmatrix},$$

which we can rewrite in the form

$$H(\theta; x) = A(\theta) + \sum_{k=0,1} \sum_{\ell=1,2} \left(\sum_{i=1}^{n^{(k)}} (x_i^{(k)} - \mu)^\ell \right) \cdot A_{k,\ell}(\theta),$$

where $A(\theta) \in \mathbb{R}^{3\times 3}$ and each $A_{k,\ell}(\theta) \in \mathbb{R}^{3\times 3}$ are all continuous matrix-valued functions of θ. Therefore, we can calculate

$$H(\theta; x) - H(\theta) = \sum_{k=0,1} \left(\sum_{i=1}^{n^{(k)}} (x_i^{(k)} - \mu) \right) \cdot A_{k,1}(\theta) + \sum_{k=0,1} \left(\sum_{i=1}^{n^{(k)}} ((x_i^{(k)} - \mu)^2 - \gamma^{(k)}) \right) \cdot A_{k,2}(\theta),$$

and so

$$\|H(\theta; x) - H(\theta)\| \leq \sum_{k=0,1} \left| \sum_{i=1}^{n^{(k)}} (x_i^{(k)} - \mu) \right| \cdot \|A_{k,1}(\theta)\| + \sum_{k=0,1} \left| \sum_{i=1}^{n^{(k)}} ((x_i^{(k)} - \mu)^2 - \gamma^{(k)}) \right| \cdot \|A_{k,2}(\theta)\|.$$

Now let $r > 0$ be any constant so that $\mathbb{B}(\theta_0, r) \subseteq \Theta$, and let

$$c_r = \sup_{\theta \in \mathbb{B}(\theta_0, r)} \max_{k=0,1} \max_{\ell=1,2} \|A_{k,\ell}(\theta)\|,$$

which is finite since the $A_{k,\ell}$'s are continuous functions of θ. Then

$$\sup_{\theta\in\mathbb{B}(\theta_0,r)\cap\Theta}\|H(\theta;x)-H(\theta)\|\le c_r\left(\sum_{k=0,1}\left|\sum_{i=1}^{n^{(k)}}(x_i^{(k)}-\mu)\right|+\sum_{k=0,1}\left|\sum_{i=1}^{n^{(k)}}((x_i^{(k)}-\mu)^2-\gamma^{(k)})\right|\right).$$

By definition of the distribution of the data we see that the terms $(x_i^{(k)}-\mu)$ are independent and Gaussian, while the terms $((x_i^{(k)}-\mu)^2-\gamma^{(k)})$ are independent centered and scaled χ^2 (and therefore subexponential). An elementary calculation then verifies that

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta\in\mathbb{B}(\theta_0,r)\cap\Theta}t\cdot\|H(\theta;X)-H(\theta)\|\right\}\right]\le e^{c''t^2n}\text{ for all }|t|\le c''',$$

where c'' is chosen to be sufficiently large and $c'''>0$ is chosen to be sufficiently small. Taking $t=\frac{c^2\log n}{n}$, and choosing n sufficiently large so that $c\sqrt{\frac{\log n}{n}}\le r$ and $t\le c'''$, we have established the desired bound (C.2) with $\varepsilon_n=\frac{\log^2 n}{n}$ and c' chosen appropriately.

C.1.3 Checking Assumption 3 for Example 3

For the Gaussian spatial process (Example 3), we can calculate

$$H(\theta;x)=\frac{1}{2}x^\top\left(\frac{\partial^2}{\partial\theta^2}\Sigma_\theta^{-1}\right)x+\frac{1}{2}\frac{\partial^2}{\partial\theta^2}\log\det(\Sigma_\theta),$$

and therefore writing $\tilde{x}=\Sigma_{\theta_0}^{-1/2}x$, we have

$$\begin{aligned}
&\|H(\theta;x)-H(\theta)\|=\frac{1}{2}\left\langle xx^\top-\Sigma_{\theta_0},\frac{\partial^2}{\partial\theta^2}\Sigma_\theta^{-1}\right\rangle\\
&=\frac{1}{2}\left\langle\tilde{x}\tilde{x}^\top-\mathbf{I}_d,\Sigma_{\theta_0}^{1/2}\cdot\frac{\partial^2}{\partial\theta^2}\Sigma_\theta^{-1}\cdot\Sigma_{\theta_0}^{1/2}\right\rangle\\
&\le\frac{1}{2}\left\langle\tilde{x}\tilde{x}^\top-\mathbf{I}_d,\Sigma_{\theta_0}^{1/2}\cdot\frac{\partial^2}{\partial\theta^2}\Sigma_{\theta_0}^{-1}\cdot\Sigma_{\theta_0}^{1/2}\right\rangle+\left\|\tilde{x}\tilde{x}^\top-\mathbf{I}_d\right\|\cdot\frac{1}{2}\|\Sigma_{\theta_0}\|\left\|\frac{\partial^2}{\partial\theta^2}\Sigma_\theta^{-1}-\frac{\partial^2}{\partial\theta^2}\Sigma_{\theta_0}^{-1}\right\|\\
&\le\left\langle\tilde{x}\tilde{x}^\top-\mathbf{I}_d,\frac{1}{2}\Sigma_{\theta_0}^{1/2}\cdot\frac{\partial^2}{\partial\theta^2}\Sigma_{\theta_0}^{-1}\cdot\Sigma_{\theta_0}^{1/2}\right\rangle+\left\|\tilde{x}\tilde{x}^\top-\mathbf{I}_d\right\|\cdot\frac{1}{2}\|\Sigma_{\theta_0}\||\theta-\theta_0|\cdot\sup_{t\in[0,1]}\left\|\frac{\partial^3}{\partial\theta^3}\Sigma_{(1-t)\theta_0+t\theta}^{-1}\right\|.
\end{aligned}$$

Therefore, taking n sufficiently large so that $\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})\subseteq\Theta$,

$$\begin{aligned}
&\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})}\|H(\theta;x)-H(\theta)\|\\
&\le\left\langle\tilde{x}\tilde{x}^\top-\mathbf{I}_d,\frac{1}{2}\Sigma_{\theta_0}^{1/2}\cdot\frac{\partial^2}{\partial\theta^2}\Sigma_{\theta_0}^{-1}\cdot\Sigma_{\theta_0}^{1/2}\right\rangle+\left\|\tilde{x}\tilde{x}^\top-\mathbf{I}_d\right\|\cdot\frac{c}{2}\sqrt{\frac{\log n}{n}}\|\Sigma_{\theta_0}\|\cdot\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})}\left\|\frac{\partial^3}{\partial\theta^3}\Sigma_\theta^{-1}\right\|.
\end{aligned}$$

By [Bachoc, 2014, Proposition D.7], the eigenvalues of $\Sigma_{\theta_0}^{1/2}\cdot\frac{\partial^2}{\partial\theta^2}\Sigma_{\theta_0}^{-1}\cdot\Sigma_{\theta_0}^{1/2}$ are bounded above by a constant not depending on n, and furthermore $\|\Sigma_{\theta_0}\|$ and (for sufficiently large n)

$\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})} \left\|\frac{\partial^3}{\partial\theta^3}\Sigma_\theta^{-1}\right\|$ are bounded by constants not depending on n. Since $\tilde{x}\sim\mathcal{N}(0,\mathbf{I}_n)$, standard tail bounds on the χ^2 distribution (e.g., [Laurent and Massart, 2000, Lemma 1]) establish that

$$\mathbb{E}_{\theta_0}\left[\exp\left\{t\cdot\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})}\|H(\theta;x)-H(\theta)\|\right\}\right]\le\exp\left\{c''\cdot t^2n+t\cdot\sqrt{\frac{\log n}{n}}\cdot n\right\}\text{ for all }|t|\le c''',$$

where c'' is chosen to be sufficiently large and $c'''>0$ is chosen to be sufficiently small. Taking $t=\frac{c^2\log n}{n}$, and choosing n sufficiently large, we have established the desired bound (C.2) with $\varepsilon_n\asymp\sqrt{\frac{\log^3 n}{n}}$ and c' chosen appropriately.

C.1.4 Checking Assumption 3 for Example 4

For the multivariate t distribution (Example 4), we first note that since $\theta\in\mathbb{R}^{k\times k}$ is a matrix parameter, the Euclidean norm is given by the matrix Frobenius norm, $\|M\|_{\mathrm{F}}=\sqrt{\sum_{ij}M_{ij}^2}$. To avoid confusion, when discussing Example 4 we will write $\|M\|_{\mathrm{op}}$ for the operator norm on matrices (both for a $k\times k$ matrix, such as the parameter θ itself, or for a $k^2\times k^2$ linear operator from $\mathbb{R}^{k\times k}$ to $\mathbb{R}^{k\times k}$, such as the Hessian).

We can first calculate the Hessian, which in this setting will be a linear operator mapping from $\mathbb{R}^{k\times k}$ to $\mathbb{R}^{k\times k}$. We calculate $H(\theta;x)$ applied to any $A,B\in\mathbb{R}^{k\times k}$ as

$$\big[H(\theta;x)\big](A,B)=\frac{n}{2}\left\langle\theta^{-1/2}A\theta^{-1/2},\theta^{-1/2}B\theta^{-1/2}\right\rangle-\frac{\gamma+k}{2}\sum_{i=1}^n\frac{(x_i^\top Ax_i)\cdot(x_i^\top Bx_i)}{(\gamma+x_i^\top\theta x_i)^2}.$$

For any θ and any $a\in(0,\frac{1}{2})$, if $(1-a)\theta_0\preceq\theta\preceq(1+a)\theta_0$, we can verify that

$$(1-a)^2\cdot(\gamma+z^\top\theta z)^2\le(\gamma+z^\top\theta z)^2\le(1+a)^2\cdot(\gamma+z^\top\theta z)^2$$

for all $z\in\mathbb{R}^k$, and therefore

$$\left|\sum_{i=1}^n\frac{(x_i^\top Ax_i)\cdot(x_i^\top Bx_i)}{(\gamma+x_i^\top\theta x_i)^2}-\sum_{i=1}^n\frac{(x_i^\top Ax_i)\cdot(x_i^\top Bx_i)}{(\gamma+x_i^\top\theta_0 x_i)^2}\right|\le n\cdot\frac{2a+a^2}{1-2a}\cdot\lambda_{\min}(\theta_0)^{-2}\cdot\|A\|_{\mathrm{op}}\cdot\|B\|_{\mathrm{op}}.$$

for all A,B, where $\lambda_{\min}(\theta_0)>0$ is the minimum eigenvalue of θ_0. Since $\|\|_{\mathrm{op}}\le\|\|_{\mathrm{F}}$, we have

$$\left|\sum_{i=1}^n\frac{(x_i^\top Ax_i)\cdot(x_i^\top Bx_i)}{(\gamma+x_i^\top\theta x_i)^2}-\sum_{i=1}^n\frac{(x_i^\top Ax_i)\cdot(x_i^\top Bx_i)}{(\gamma+x_i^\top\theta_0 x_i)^2}\right|\le n\cdot\frac{2a+a^2}{1-2a}\cdot\lambda_{\min}(\theta_0)^{-2}$$

for all A,B with $\|A\|_{\mathrm{F}},\|B\|_{\mathrm{F}}\le1$. This is sufficient to verify that

$$\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})}\|H(\theta;x)-H(\theta)\|_{\mathrm{op}}\le\|H(\theta_0;x)-H(\theta_0)\|_{\mathrm{op}}+\sqrt{n\log n}\cdot3c\lambda_{\min}(\theta_0)^{-2}$$

for all sufficiently large n. Therefore, for sufficiently large n,

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta\in\mathbb{B}(\theta_0,c\sqrt{\frac{\log n}{n}})\cap\Theta}\frac{c^2\log n}{n}\cdot\|H(\theta;X)-H(\theta)\|_{\mathrm{op}}\right\}\right]$$
$$\leq \exp\left\{\frac{c^2\log n}{n}\cdot\sqrt{n\log n}\cdot 3c\lambda_{\min}(\theta_0)^{-2}\right\}\cdot\mathbb{E}_{\theta_0}\left[\exp\left\{\frac{c^2\log n}{n}\cdot\|H(\theta_0;X)-H(\theta_0)\|_{\mathrm{op}}\right\}\right].$$

Next, $H(\theta_0;X)$ is equal to a constant plus a sum of n i.i.d. terms, with each term bounded uniformly, since

$$\left|\frac{\gamma+k}{2}\frac{(X_i^\top AX_i)\cdot(X_i^\top BX_i)}{(\gamma+X_i^\top\theta_0X_i)^2}\right|\leq\frac{\gamma+k}{2}\cdot\lambda_{\min}(\theta_0)^{-2}$$

holds for all A,B with $\|A\|_{\mathrm{F}},\|B\|_{\mathrm{F}}\leq 1$, almost surely over X_i. Therefore, by the matrix Hoeffding inequality [Tropp, 2012, Theorem 1.3], we have

$$\mathbb{P}_{\theta_0}\left(n^{-1/2}\|H(\theta_0;X)-H(\theta_0)\|_{\mathrm{op}}>t\right)\leq 2k^2\exp\left\{-\frac{t^2}{8\cdot(\frac{\gamma+k}{2})^2\cdot\lambda_{\min}(\theta_0)^{-4}}\right\}$$

for any $t>0$. In other words, $n^{-1/2}\|H(\theta_0;X)-H(\theta_0)\|_{\mathrm{op}}$ is subgaussian with parameter not depending on n and therefore

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\frac{c^2\log n}{n}\cdot\|H(\theta_0;X)-H(\theta_0)\|_{\mathrm{op}}\right\}\right]\leq\exp\left\{\frac{c''\log^2 n}{n}\right\}$$

for an appropriately chosen c''.

Combining everything, we have established that the bound (C.2) holds with $\varepsilon_n\asymp\sqrt{\frac{\log^3 n}{n}}$ and c' chosen appropriately.

C.2 Checking Assumption 2

Before giving proofs for our specific examples, we pause to discuss Assumption 2 more generally, to see that this assumption will be plausible for many common settings (beyond the few that we study here). We consider the following general scenario. Suppose that we have access to a consistent initial estimate $\widehat{\theta}_{\mathrm{init}}(X)$ of θ_0. Then under some standard conditions on the negative log-likelihood surface, by constraining $\widehat{\theta}(X,W)$ to a neighborhood of $\widehat{\theta}_{\mathrm{init}}(X)$, we can ensure that $\widehat{\theta}(X,W)$ will satisfy the needed assumptions.

Lemma 6. *Let*

$$\widehat{\theta}_{\mathrm{init}}:\mathcal{X}\to\Theta\quad\text{and}\quad\widehat{r}_{\mathrm{init}}:\mathcal{X}\to\mathbb{R}_+$$

be any maps such that $\mathbb{B}(\widehat{\theta}_{\mathrm{init}}(x),\widehat{r}_{\mathrm{init}}(x))\subseteq\Theta$ *for all* $x\in\mathcal{X}$. *Suppose that, under the distribution* $X\sim P_{\theta_0}$, *the following statements all hold with probability at least* $1-\delta_{\mathrm{init}}(\theta_0)$:

$$\begin{cases}\|\widehat{\theta}_{\mathrm{init}}(X)-\theta_0\|\leq r_{\mathrm{init}}(\theta_0),\\ \mathcal{L}(\theta;X)\text{ has a FOSP in }\mathbb{B}(\theta_0,r_{\mathrm{init}}(\theta_0)),\\ \nabla^2_\theta\mathcal{L}(\theta;X)\succeq\lambda_{\mathrm{cvx}}(\theta_0)\mathbf{I}_d\text{ for all }\theta\in\mathbb{B}(\theta_0,r_{\mathrm{cvx}}(\theta_0)),\\ 3r_{\mathrm{init}}(\theta_0)\leq\widehat{r}_{\mathrm{init}}(X)\leq r_{\mathrm{cvx}}(\theta_0)-r_{\mathrm{init}}(\theta_0),\end{cases}\tag{C.3}$$

for some constants $r_{\text{init}}(\theta_0), r_{\text{cvx}}(\theta_0), \lambda_{\text{cvx}}(\theta_0) > 0$. *If* $\widehat{\theta} : \mathcal{X} \times \mathbb{R}^d \mapsto \Theta$ *is any function that maps each point* (x, w) *to some FOSP of the constrained optimization problem*

$$\underset{\theta \in \mathbb{B}(\widehat{\theta}_{\text{init}}(x), r(\widehat{\theta}_{\text{init}}(x)))}{\arg\min} \quad \mathcal{L}(\theta; x, w),$$

then Assumption 2 is satisfied with

$$r(\theta_0) = 2r_{\text{init}}(\theta_0) \text{ and } \delta(\theta_0) = \delta_{\text{init}}(\theta_0) + \exp\left\{-\frac{1}{2}\max\left\{\frac{r_{\text{init}}(\theta_0)\lambda_{\text{cvx}}(\theta_0)}{\sigma} - 1, 0\right\}^2\right\}.$$

With this lemma in place, we will now turn to verifying that its conditions hold for each of our four examples. Specifically, for each example, we will propose an initial estimator $\widehat{\theta}_{\text{init}}(X)$ such that the conditions of the lemma are satisfied with $r_{\text{init}}(\theta_0) \asymp \sqrt{\frac{\log n}{n}}$ and $r_{\text{cvx}}(\theta_0) \asymp 1$ and $\lambda_{\text{cvx}}(\theta_0) \asymp n$.

C.2.1 Checking the conditions of Lemma 6: general recipe

After fixing some $\theta_0 \in \Theta$, each proof will follow the same general recipe:

- We will verify that
$$H(\theta_0) \succeq C_1 n \mathbf{I}_d, \tag{C.4}$$
where $C_1 > 0$ does not depend on n. Combined with Assumption 3 (which we verified above for each of our examples), this means that for sufficiently large n it holds that $\nabla^2_\theta \mathcal{L}(\theta; X) = H(\theta; X) \succeq C_2 n \mathbf{I}_3$ for all $\theta \in \mathbb{B}(\theta_0, C_3)$ for appropriately chosen $C_2, C_3 > 0$, with probability at least $1 - n^{-1}$. Thus we can take $\lambda_{\text{cvx}} = C_2$ and $r_{\text{cvx}} = C_3$.

- We will define an initial estimator $\widehat{\theta}_{\text{init}}(x)$ and will prove that we can find a constant C_4 not depending on n such that
$$\mathbb{P}_{\theta_0}\left(\|\widehat{\theta}_{\text{init}}(X) - \theta_0\| \le C_4\sqrt{\frac{\log n}{n}}\right) \ge 1 - n^{-1} \tag{C.5}$$
for all sufficiently large n. Thus we can take $r_{\text{init}} = C_4\sqrt{\frac{\log n}{n}}$. Furthermore, choosing $\widehat{r}_{\text{init}}(x)$ to be any function of n that vanishes slower than $\sqrt{\frac{\log n}{n}}$ (e.g., $\widehat{r}_{\text{init}}(x) \equiv n^{-1/4}$), we have verified that $3r_{\text{init}}(\theta_0) \le \widehat{r}_{\text{init}}(X) \le r_{\text{cvx}}(\theta_0) - r_{\text{init}}(\theta_0)$ holds.

- Finally we will show that we can find a constant C_5 not depending on n such that
$$\mathbb{P}_{\theta_0}\left(\|\nabla_\theta \log f(X; \theta_0)\| \le C_5\sqrt{n \log n}\right) \ge 1 - n^{-1}, \tag{C.6}$$
for all sufficiently large n. Combined with the bound $\nabla^2_\theta \mathcal{L}(\theta; X) \succeq C_2 n \mathbf{I}_3$ for all $\theta \in \mathbb{B}(\theta_0, C_3)$ that is already established, this means that $\mathcal{L}(\theta; X) = -\log f(X; \theta)$ has a FOSP in $\mathbb{B}(\theta_0, C_5 C_2^{-1}\sqrt{\frac{\log n}{n}})$ and so we can take $r_{\text{init}}(\theta_0) = C_5 C_2^{-1}\sqrt{\frac{\log n}{n}}$.

C.2.2 Checking the conditions of Lemma 6 for Example 1

For the canonical GLM setting (Example 1), first we have

$$H(\theta_0; x) = \sum_{i=1}^{n} Z_i Z_i^\top \cdot a''(Z_i^\top \theta_0) \succeq n \cdot C_1 \cdot \mathbf{I}$$

for some $C_1 > 0$ that does not depend on n, since we have assumed $\max_{ij} |Z_{ij}|$ is bounded by a constant and $\frac{1}{n}\sum_i Z_i Z_i^\top \succeq \lambda_0 \mathbf{I}_d$. Thus (C.4) holds. Next we verify (C.5). Since the negative log-likelihood is strictly convex everywhere, we can define $\widehat{\theta}_{\text{init}}(x)$ to equal a global minimizer of $-\log f(x;\theta)$, if one exists (i.e., finding a global minimizer is computationally feasible since it is a differentiable and strictly convex minimization problem). Therefore, if a FOSP exists in a $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$ neighborhood of θ_0 (as we will establish next), then (C.5) is satisfied. Finally we check (C.6) to verify the existence of the FOSP. We calculate

$$\nabla_\theta[-\log f(x;\theta_0)] = \sum_{i=1}^{n} Z_i \left(a'(Z_i^\top \theta_0) - x_i\right),$$

and by standard calculations for GLMs, X is subexponential with

$$\mathbb{E}_{\theta_0}\left[e^{tX_i}\right] = e^{a(Z_i^\top \theta_0 + t) - a(Z_i^\top \theta_0)}$$

for any $t \in \mathbb{R}$ and for each $i = 1, \ldots, n$. Since we have assumed $\max_{ij} |Z_{ij}|$ is bounded by a constant, proving (C.6) is a standard calculation.

C.2.3 Checking the conditions of Lemma 6 for Example 2

For the Behrens–Fisher problem (Example 2), write $\theta_0 = (\mu_0, \gamma_0^{(0)}, \gamma_0^{(1)})$. We first calculate

$$H(\theta_0) = \mathbb{E}_{\theta_0}\left[H(\theta_0; X)\right] = \begin{pmatrix} \frac{n^{(0)}}{\gamma_0^{(0)}} + \frac{n^{(1)}}{\gamma^{(1)}}_0 & 0 & 0 \\ 0 & \frac{n^{(0)}}{2(\gamma_0^{(0)})^2} & 0 \\ 0 & 0 & \frac{n^{(1)}}{2(\gamma_0^{(1)})^2} \end{pmatrix} \succeq cn \cdot \frac{\min\{n^{(0)}, n^{(1)}\}}{\max\{n^{(0)}, n^{(1)}\}} \cdot \mathbf{I}_3,$$

where the inequality holds for some $c > 0$ depending only on θ_0. Recalling that we have assumed $\frac{\max\{n^{(0)}, n^{(1)}\}}{\min\{n^{(0)}, n^{(1)}\}}$ is bounded by a constant, this means that

$$H(\theta_0) \succeq c' n \mathbf{I}_3$$

for some $c' > 0$ that does not depend on n, which verifies (C.4).

Next we define an initial estimator

$$\widehat{\theta}_{\text{init}}(x) = \left(\widehat{\mu}_{\text{init}}(x), \widehat{\gamma}^{(0)}_{\text{init}}(x), \widehat{\gamma}^{(1)}_{\text{init}}(x)\right)$$

where

$$\widehat{\mu}_{\text{init}}(x) = \frac{1}{n}\left(\sum_{i=1}^{n^{(0)}} x_i^{(0)} + \sum_{i=1}^{n^{(1)}} x_i^{(1)}\right)$$

for $n = n^{(0)} + n^{(1)}$, and

$$\widehat{\gamma}_{\text{init}}^{(k)}(x) = \frac{1}{n^{(k)}} \sum_{i=1}^{n^{(k)}} (x_i^{(k)} - \widehat{\mu}_{\text{init}}(x))^2$$

for each $k = 0, 1$. By standard Gaussian and χ^2 tail bounds, we can easily see that for sufficiently large c'' (not depending on n) it holds that

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\theta}_{\text{init}}(X) - \theta_0\| \le c'' \sqrt{\frac{\log n}{n}} \right) \ge 1 - n^{-1}$$

for sufficiently large n, which verifies (C.5).

Finally, we calculate

$$\nabla \log f(x;\theta) = - \begin{pmatrix} -\sum_{k=0,1} \frac{\sum_{i=1}^{n^{(k)}} (x_i^{(k)} - \mu)}{\gamma^{(k)}} \\ \frac{n^{(0)}}{2\gamma_0^{(0)}} - \sum_{i=1}^{n^{(0)}} \frac{(x_i^{(0)} - \mu_0)^2}{2(\gamma_0^{(0)})^2} \\ \frac{n^{(1)}}{2\gamma_0^{(1)}} - \sum_{i=1}^{n^{(1)}} \frac{(x_i^{(1)} - \mu_0)^2}{2(\gamma_0^{(1)})^2} \end{pmatrix},$$

and therefore each entry of $\nabla \log f(X;\theta_0)$ is a sum of n or $n^{(0)}$ or $n^{(1)}$ many i.i.d. zero-mean subexponential terms. Therefore, we can find a constant c''' such that

$$\mathbb{P}_{\theta_0}\left(\|\nabla \log f(X;\theta_0)\| \le c''' \sqrt{n \log n} \right) \ge 1 - n^{-1}$$

for sufficiently large n, which verifies (C.6) and thus completes the proof.

C.2.4 Checking the conditions of Lemma 6 for Example 3

For the Gaussian spatial process (Example 3), first, recall our calculation

$$H(\theta; x) = \frac{1}{2} x^\top \left(\frac{\partial^2}{\partial \theta^2} \Sigma_{\theta_0}^{-1} \right) x + \frac{1}{2} \frac{\partial^2}{\partial \theta^2} \log\det(\Sigma_{\theta_0}),$$

which we can calculate explicitly as

$$\begin{aligned} H(\theta; X) = \frac{1}{2} x^\top \left(-\Sigma_\theta^{-1}(D \circ D \circ \Sigma_\theta)\Sigma_\theta^{-1} + 2\Sigma_\theta^{-1}(D \circ \Sigma_\theta)\Sigma_\theta^{-1}(D \circ \Sigma_\theta)\Sigma_\theta^{-1} \right) x \\ + \frac{1}{2}\text{trace}(\Sigma_\theta^{-1/2}(D \circ D \circ \Sigma_\theta)\Sigma_\theta^{-1/2}) - \frac{1}{2}\|\Sigma_\theta^{-1/2}(D \circ \Sigma_\theta)\Sigma_\theta^{-1/2}\|^2, \end{aligned}$$

and so since $\mathbb{E}_{\theta_0}\left[XX^\top\right] = \Sigma_{\theta_0}$, we have

$$H(\theta_0) = \mathbb{E}_{\theta_0}\left[H(\theta_0; X)\right] = \frac{1}{2}\|\Sigma_{\theta_0}^{-1/2}(D \circ \Sigma_{\theta_0})\Sigma_{\theta_0}^{-1/2}\|^2 \ge \frac{1}{2}\lambda_{\min}(\Sigma_{\theta_0})^{-2}\|D \circ \Sigma_{\theta_0}\|^2.$$

We know from [Bachoc, 2014, Proposition D.7] that Σ_{θ_0} has eigenvalues bounded above and below by positive constants. Furthermore,

$$\|D \circ \Sigma_{\theta_0}\|^2 = \sum_{i=1}^n \sum_{j=1}^n D_{ij}^2 \cdot (\Sigma_{\theta_0})_{ij}^2 \ge \sum_{(i,j)\in E} D_{ij}^2 \cdot (\Sigma_{\theta_0})_{ij}^2 = \sum_{(i,j)\in E} 1 \cdot e^{-2\theta_0} = e^{-2\theta_0} \cdot |E|,$$

where $E \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ be the set of all pairs (i,j) such that $D_{ij} = 1$. Since $|E| \geq n$, we have shown that (C.4) holds for some appropriately chosen C_1 that does not depend on n.

Next we need to define our initial estimator to satisfy (C.5). We will define a simple choice for intuition (this choice is of course not necessarily optimal in any sense). Define

$$\widehat{\theta}_{\text{init}}(x) = -\log\left(\frac{1}{|E|}\sum_{(i,j)\in E} x_i x_j\right).$$

We need to check that, with probability at least $1-n^{-1}$, $|\widehat{\theta}_{\text{init}}(X) - \theta_0| \leq C_4\sqrt{\frac{\log n}{n}}$ for some constant C_4 not depending on n. Since $\theta_0 > 0$, it is equivalent to check that, with probability at least $1-n^{-1}$,

$$\left|\frac{1}{|E|}\sum_{(i,j)\in E} X_i X_j - e^{-\theta_0}\right| \leq C'\sqrt{\frac{\log n}{n}}$$

for some constant C' not depending on n. Let A be the adjacency matrix, with entry $A_{ij} = \mathbb{1}\{(i,j)\in E\}$, and let $U\Lambda U^\top = \Sigma_{\theta_0}^{1/2} A \Sigma_{\theta_0}^{1/2}$ be an eigendecomposition. Then

$$\left|\frac{1}{|E|}\sum_{(i,j)\in E} X_i X_j - e^{-\theta_0}\right| = \frac{1}{|E|}\left|\left\langle XX^\top - \Sigma_{\theta_0}, A\right\rangle\right| = \frac{1}{|E|}\left|\left\langle (U^\top\Sigma_{\theta_0}^{-1/2}X)(U^\top\Sigma_{\theta_0}^{-1/2}X)^\top - \mathbf{I}_n, \Lambda\right\rangle\right|.$$

Since $U^\top\Sigma_{\theta_0}^{-1/2}X \sim \mathcal{N}(0,\mathbf{I}_n)$, while $|E| \geq n$, the desired bound holds as long as the values $\Lambda_{11},\dots,\Lambda_{nn}$ (i.e., the eigenvalues of $\Sigma_{\theta_0}^{1/2}A\Sigma_{\theta_0}^{1/2}$) are bounded by some constant C'' not depending on n. Since the eigenvalues of Σ_{θ_0} are bounded by a constant not depending on n by [Bachoc, 2014, Proposition D.7], equivalently we need to verify that $\|A\| \leq C'''$ for some constant C''' not depending on n—in fact, since A is the adjacency matrix of a graph where each vertex has at most $2k$ many neighbors, we have $\|A\| \leq 2k$. This establishes (C.5).

Finally we verify (C.6). We calculate

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(x;\theta_0) &= -\frac{1}{2}x^\top\left(\frac{\partial}{\partial\theta}\Sigma_{\theta_0}^{-1}\right)x - \frac{1}{2}\frac{\partial}{\partial\theta}\log\det(\Sigma_{\theta_0}) \\ &= -\frac{1}{2}(\Sigma_{\theta_0}^{-1/2}x)^\top\cdot\Sigma_{\theta_0}^{1/2}\left(\frac{\partial}{\partial\theta}\Sigma_{\theta_0}^{-1}\right)\Sigma_{\theta_0}^{1/2}\cdot(\Sigma_{\theta_0}^{-1/2}x) - \frac{1}{2}\frac{\partial}{\partial\theta}\log\det(\Sigma_{\theta_0})\end{aligned}$$

We know that $\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X;\theta_0)\right] = 0$, and moreover, $\Sigma_{\theta_0}^{-1/2}X \sim \mathcal{N}(0,\mathbf{I}_n)$ and so this quantity has distribution equal to a weighted sum of centered χ^2 random variables. By [Bachoc, 2014, Proposition D.7] we know that the eigenvalues of the matrix $\Sigma_{\theta_0}^{1/2}\left(\frac{\partial}{\partial\theta}\Sigma_{\theta_0}^{-1}\right)\Sigma_{\theta_0}^{1/2}$ are bounded by a constant that does not depend on n, standard χ^2 tail bounds (see, e.g., [Laurent and Massart, 2000, Lemma 1]) establish that (C.6) holds for an appropriately chosen C_5 not depending on n.

C.2.5 Checking the conditions of Lemma 6 for Example 4

For the multivariate t distribution (Example 4), calculations in [Lange et al., 1989, Appendix B] show that

$$\big[H(\theta_0)\big](M,M) = \frac{n}{2}\left(\frac{\gamma+k}{\gamma+k+2}\|\theta_0^{-1/2}M\theta_0^{-1/2}\|_{\mathrm{F}}^2 - \frac{1}{\gamma+k+2}\mathrm{trace}(\theta_0^{-1/2}M\theta_0^{-1/2})^2\right).$$

Since $\mathrm{trace}(A) \le \sqrt{k}\|A\|_{\mathrm{F}}$ for any $A \in \mathbb{R}^{k\times k}$, then we have

$$\big[H(\theta_0)\big](M,M) \ge \frac{n}{2}\cdot\frac{\gamma}{\gamma+k+2}\|\theta_0^{-1/2}M\theta_0^{-1/2}\|_{\mathrm{F}}^2,$$

and so (C.4) holds with $C_1 = \frac{1}{2}\cdot\frac{\gamma}{\gamma+k+2}\cdot\|\theta_0\|_{\mathrm{op}}^{-2}$.

Next, we define our initial estimator. We will work with the Kendall's τ correlation: given a data point $x \in (\mathbb{R}^k)^n$, for each $j, j' \in \{1,\dots,k\}$ define

$$T_{jj'} = \frac{1}{\binom{n}{2}}\sum_{1\le i<i'\le n}\mathrm{sign}\left((x_{ij}-x_{i'j})\cdot(x_{ij'}-x_{i'j'})\right),$$

let $S = S(x) \in \mathbb{R}^{k\times k}$ be defined with entries $S_{jj'} = \sin\left(\frac{\pi}{2}\cdot T_{jj'}\right)$. It is well known that for a continuous elliptical distribution (such as the multivariate t), this transformation yields an unbiased estimate of the correlation matrix. We will also estimate $V_j = \frac{\text{Median of } |x_{1j}|,\dots,|x_{nj}|}{\text{0.75-quantile of } t_\gamma}$, and let $\Sigma = \Sigma(x)$ have entries

$$\Sigma_{jj'} = S_{jj'}\sqrt{V_jV_{j'}}.$$

Next let $\widehat{\theta}_{\mathrm{init}}(x) = \Sigma(x)^{-1}$ (or define it to take any value if $\Sigma(x)$ is not invertible).

By [Barber and Kolar, 2018, Corollary 4.8], if $n \ge k\log n$, then with probability at least $1-n^{-1}$,

$$\|S(X)-S_*\|_{\mathrm{op}} \le C\sqrt{\frac{k\log n}{n}}$$

for a universal constant C, where S_* is the true correlation matrix, i.e.,

$$(S_*)_{jj'} = \frac{(\theta_0^{-1})_{jj'}}{\sqrt{(\theta_0)_{jj}^{-1}(\theta_0^{-1})_{j'j'}}}.$$

We also have $X_{ij} \overset{\mathrm{iid}}{\sim} \sqrt{(\theta_0^{-1})_{jj}}\cdot t_\gamma$ (a univariate t distribution) and so we can easily verify that

$$\max_{j=1,\dots,k}|V_j - (\theta_0^{-1})_{jj}| \le C'\sqrt{\frac{\log n}{n}}$$

with probability at least $1-n^{-1}$. Combining these bounds, this means that

$$\|\widehat{\theta}_{\mathrm{init}}(X)-\theta_0\|_{\mathrm{op}} \le C''\sqrt{\frac{k\log n}{n}}.$$

Since this is a $k \times k$ matrix, therefore

$$\|\widehat{\theta}_{\text{init}}(X) - \theta_0\|_{\text{F}} \leq C'' \sqrt{\frac{k^2 \log n}{n}} = C''' \sqrt{\frac{\log n}{n}},$$

which verifies (C.5).

Finally we check (C.6). We calculate

$$\nabla \log f(X; \theta_0) = \frac{n}{2} \theta_0^{-1} - \frac{\gamma + k}{2} \sum_{i=1}^{n} \frac{X_i X_i^\top}{\gamma + X_i^\top \theta_0 X_i},$$

which is a sum of n i.i.d. mean-zero terms. Observe also that for any $z \in \mathbb{R}^k$, $\|\frac{zz^\top}{\gamma + z^\top \theta_0 z}\|_{\text{op}} \leq \lambda_{\min}(\theta_0)^{-1}$, so the terms are uniformly bounded. By the matrix Hoeffding inequality [Tropp, 2012, Theorem 1.3] along with the bound $\|\nabla \log f(X; \theta_0)\|_{\text{F}} \leq \sqrt{k} \|\nabla \log f(X; \theta_0)\|_{\text{op}}$, we therefore have

$$\mathbb{P}_{\theta_0} \left(\|\nabla \mathcal{L}(\theta_0; X)\|_{\text{F}} \geq t\sqrt{k} \right) \leq 2k \exp \left\{ - \frac{t^2}{8n \cdot (\frac{\gamma + k}{2})^2 \cdot \lambda_{\min}(\theta_0)^{-2}} \right\}$$

for any $t > 0$. Taking $t \asymp \sqrt{n \log n}$ is sufficient to establish (C.6).

C.3 Proof of Lemma 6

Suppose that the statements (C.3) all hold, which is satisfied with probability at least $1 - \delta_{\text{init}}(\theta_0)$ by assumption. Suppose also that the random vector W satisfies

$$\|W\| < \frac{r_{\text{init}}(\theta_0) \lambda_{\text{cvx}}(\theta_0)}{\sigma}.$$

Since $\|W\|^2 \sim \frac{1}{d} \chi^2_d$ by definition, using standard χ^2 tail bounds (see, e.g., [Laurent and Massart, 2000, Lemma 1]) we can calculate

$$\mathbb{P} \left(\|W\| < \frac{r_{\text{init}}(\theta_0) \lambda_{\text{cvx}}(\theta_0)}{\sigma} \right) \geq 1 - \exp \left\{ -\frac{1}{2} \max \left\{ \frac{r_{\text{init}}(\theta_0) \lambda_{\text{cvx}}(\theta_0)}{\sigma} - 1, 0 \right\}^2 \right\}.$$

Therefore, with probability at least $1 - \delta(\theta_0)$ (where $\delta(\theta_0)$ is defined as in the statement of the lemma), the bounds (C.3) all hold and $\|W\|$ satisfies the bound above. From this point on we will assume these bounds all hold.

Let $\theta_* \in \mathbb{B}(\theta_0, r_{\text{init}}(\theta_0))$ be a FOSP of $\mathcal{L}(\theta; X)$, and let $\widehat{\theta} = \widehat{\theta}(X, W) \in \mathbb{B}(\widehat{\theta}_{\text{init}}(X), \widehat{r}_{\text{init}}(X))$ be a FOSP of the constrained problem

$$\min_{\theta \in \mathbb{B}(\widehat{\theta}_{\text{init}}(X), \widehat{r}_{\text{init}}(X))} \mathcal{L}(\theta; X, W).$$

Then

$$\|\theta_* - \widehat{\theta}_{\text{init}}(X)\| \leq \|\widehat{\theta}_{\text{init}}(X) - \theta_0\| + \|\theta_* - \theta_0\| \leq 2 r_{\text{init}}(\theta_0) \leq \widehat{r}_{\text{init}}(X),$$

and so θ_* also lies in the convex constraint set $\mathbb{B}(\widehat{\theta}_{\text{init}}(X), \widehat{r}_{\text{init}}(X))$. Since $\|\widehat{\theta}_{\text{init}}(X) - \theta_0\| \leq r_{\text{init}}(\theta_0) \leq r_{\text{cvx}}(\theta_0) - \widehat{r}_{\text{init}}(X)$, this means that $\widehat{\theta}$ and θ_* both lie in $\mathbb{B}(\theta_0, r_{\text{cvx}}(\theta_0))$, and so we have $\lambda_{\text{cvx}}(\theta_0)$-strong convexity in this region. Therefore we have

$$\begin{aligned}
0 &\leq (\theta_* - \widehat{\theta})^\top \nabla_\theta \mathcal{L}(\widehat{\theta}; X, W) \\
&= (\theta_* - \widehat{\theta})^\top \nabla_\theta \mathcal{L}(\widehat{\theta}; X) + \sigma(\theta_* - \widehat{\theta})^\top W \\
&\leq (\theta_* - \widehat{\theta})^\top \nabla_\theta \mathcal{L}(\theta_*; X) - \lambda_{\text{cvx}}(\theta_0)\|\theta_* - \widehat{\theta}\|^2 + \sigma\|\theta_* - \widehat{\theta}\|\|W\| \\
&\leq -\lambda_{\text{cvx}}(\theta_0)\|\theta_* - \widehat{\theta}\|^2 + \sigma\|\theta_* - \widehat{\theta}\|\|W\| \\
&< -\lambda_{\text{cvx}}(\theta_0)\|\theta_* - \widehat{\theta}\|^2 + r_{\text{init}}(\theta_0)\lambda_{\text{cvx}}(\theta_0)\|\theta_* - \widehat{\theta}\|,
\end{aligned}$$

where the next-to-last step holds since θ_* is a FOSP of the unconstrained problem $\min_\theta \mathcal{L}(\theta; X)$, and the last step holds as long as $\|\theta_* - \widehat{\theta}\| > 0$ by our bound on $\|W\|$. Therefore, we must have

$$\|\widehat{\theta} - \theta_*\| < r_{\text{init}}(\theta_0).$$

In particular this implies

$$\|\widehat{\theta} - \theta_0\| \leq \|\widehat{\theta} - \theta_*\| + \|\theta_* - \theta_0\| < 2r_{\text{init}}(\theta_0) \leq \widehat{r}_{\text{init}}(X).$$

Finally, we need to check that $\widehat{\theta}$ is a SSOSP. We have

$$\|\widehat{\theta} - \widehat{\theta}_{\text{init}}(X)\| \leq \|\widehat{\theta} - \theta_0\| + \|\widehat{\theta}_{\text{init}}(X) - \theta_0\| < 3r_{\text{init}}(\theta_0) \leq \widehat{r}_{\text{init}}(X),$$

which means that $\widehat{\theta}$ is in the interior of the constraint set $\mathbb{B}(\widehat{\theta}_{\text{init}}(X), \widehat{r}_{\text{init}}(X))$. Therefore, $\widehat{\theta}$ must be a FOSP of the unconstrained problem $\min_\theta \mathcal{L}(\theta; X, W)$. Finally, since $\widehat{\theta} \in \mathbb{B}(\theta_0, r_{\text{cvx}}(\theta_0))$ as calculated above, $\mathcal{L}(\theta; X)$ (and therefore also $\mathcal{L}(\theta; X, W)$) has strong convexity at $\theta = \widehat{\theta}$. This completes the proof.

D Computational considerations

D.1 Optimization of (2.2)

If the unperturbed penalized maximum likelihood problem is (strongly) convex, then (2.2) is also (strongly) convex. Since the linear perturbation only changes the gradient by a fixed constant and does not affect the Hessian, any convex solver that relies on first- and second-order derivatives to solve the unperturbed problem can be immediately adapted to run on (2.2). Note that even strong convexity does not guarantee the unperturbed penalized maximum likelihood problem has any local optima, since Θ could be constrained to a region with no minima. However, as long as the unperturbed problem is convex and has a local optimum, the perturbation can only lead to a lack of local optima if there exists a direction $z \in \mathbb{R}^d$ such that

$$-\sigma W^\top z \geq \max_{\theta \in \Theta} \{\nabla_\theta \mathcal{L}(\theta; X)^\top z\}. \tag{D.1}$$

Since we control σ in the aCSS algorithm, we can always choose it to be sufficiently small as to make (D.1) very unlikely (and moreover, if $\Theta = \mathbb{R}^d$ is unconstrained and the unperturbed

problem is strongly convex, then (D.1) cannot occur at any σ). Indeed, when X is composed of n i.i.d. samples, the right-hand side of (D.1) will grow at a rate of $\sqrt{n}$, while in Section 3.2, we noted that Theorem 1 required σ, and hence the left-hand side of (D.1), to grow at a rate that is vanishing compared to $\sqrt{n}$. The same story holds for non-convex functions locally for a well-behaved basin of attraction: the random perturbation can cause problems but not if you choose it sufficiently small. Note that the cost of $\widehat{\theta}$ failing to return a SSOSP of (2.2) is conservativeness of the aCSS test (but not loss of validity!), since when it fails to return a SSOSP the test will return a p-value of 1.

Although σ can always be chosen to be very small, this can incur a different computational cost in terms of sampling the copies $\widetilde{X}^{(m)}$. In particular, as we will see in the next subsection, reducing σ leads to "smaller" MCMC steps, i.e., starting at some state X' and taking a single step in the reversible Markov chain we will use for sampling will produce a state that is highly-related to X' or may even be identical to it with high probability. One solution to this is to simply take L, the number of steps we take in the Markov chain between samples, to be very large, so at least with sufficient computational resources it should always be possible to choose σ sufficiently small so as to not adversely affect the optimization of (2.2) relative to the unperturbed maximum likelihood problem.

D.2 Sampling the conditional randomizations

Due to the conditioning on $\widehat{\theta}$, the solution to an optimization problem, we only expect to be able to perform the exact sampling i.i.d. from Equation (2.7) in special cases when both the conditional distribution of X is very simple and $\widehat{\theta}$ can be found in closed form. Aside from very special cases, we expect almost any model and/or estimator to require one of the MCMC samplers.

Recall the density we are targeting in Equation (2.7):

$$p_{\widehat{\theta}}(x \,|\, \widehat{\theta}) \propto f(x; \widehat{\theta}) \cdot \exp\left\{ -\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|^2}{2\sigma^2/d} \right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x) \right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}},$$

with respect to the base measure $\nu_{\mathcal{X}}$. Both MCMC sampling schemes assume the ability to take steps in a reversible Markov chain whose stationary distribution has the above density. We will now show that it is feasible to construct an efficient sampling scheme using Metropolis–Hastings (MH).

Given $\widehat{\theta}$, we first choose a proposal distribution $q_{\widehat{\theta}}(x \,|\, x')$—we will discuss this choice below. Fixing $q_{\widehat{\theta}}(x \,|\, x')$, we can write the MH acceptance probability for a proposal x from a previous iteration x' as

$$A_{\widehat{\theta}}(x \,|\, x') := \min\left\{ 1, \frac{p_{\widehat{\theta}}(x \,|\, \widehat{\theta})\, q_{\widehat{\theta}}(x' \,|\, x)}{p_{\widehat{\theta}}(x' \,|\, \widehat{\theta})\, q_{\widehat{\theta}}(x \,|\, x')} \right\}.$$

Our reversible MCMC is then given by the following:

- Starting at state x', generate a proposal x according to the proposal distribution $q_{\widehat{\theta}}(\cdot \,|\, x')$.

- With probability $A_{\widehat{\theta}}(x \,|\, x')$, set the next state to equal x. Otherwise, the next state is set to equal x'.

To verify that this yields a computationally feasible method, we need to check two things: first, that the acceptance probability $A_{\widehat{\theta}}(x \,|\, x')$ is not too low (i.e., its average value is bounded away from zero), in order to ensure that our chain length L does not need to be taken to be too large, and second, that the acceptance probability $A_{\widehat{\theta}}(x \,|\, x')$ can be calculated efficiently. The first consideration, ensuring that $A_{\widehat{\theta}}(x \,|\, x')$ is not too low, will be specific to the problem and will discuss this for specific examples below. To check that we can efficiently calculate the acceptance probability $A_{\widehat{\theta}}(x \,|\, x')$, by definition of $p_{\widehat{\theta}}(\cdot \,|\, \widehat{\theta})$ we see that $A_{\widehat{\theta}}(x \,|\, x')$ can be written as

$$A_{\widehat{\theta}}(x \,|\, x') = \min\left\{1,\ \frac{q_{\widehat{\theta}}(x' \,|\, x)}{q_{\widehat{\theta}}(x \,|\, x')} \cdot \frac{f(x;\widehat{\theta}) \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta};x)\|^2}{2\sigma^2/d}\right\} \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta};x)\right)}{f(x';\widehat{\theta}) \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta};x')\|^2}{2\sigma^2/d}\right\} \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta};x')\right)} \cdot \frac{\mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}}{\mathbb{1}_{x' \in \mathcal{X}_{\widehat{\theta}}}}\right\}.$$

We consider the three fractions appearing in this expression. The first two are generally straightforward to calculate, but the last ratio, with the indicator variables, requires more careful consideration. In the denominator, we will have $\mathbb{1}_{x' \in \mathcal{X}_{\widehat{\theta}}} = 1$ always, since x' denotes the current state which is therefore a draw from the density (2.7) supported on $\mathcal{X}_{\widehat{\theta}}$. Turning to the numerator, however, we see that we do need to verify that our proposed state x also lies in $\mathcal{X}_{\widehat{\theta}}$. To do so, we observe that for any θ,

$$\begin{aligned}
\mathbb{1}_{x \in \mathcal{X}_\theta} &= \mathbb{1}\left\{\text{for some } w \in \mathbb{R}^d,\ \widehat{\theta}(x,w) = \theta \text{ and } \theta \text{ is a SSOSP of } \mathcal{L}(\theta;x,w)\right\} \\
&= \mathbb{1}\left\{\widehat{\theta}\left(x, -\frac{\nabla_\theta \mathcal{L}(\theta;x)}{\sigma}\right) = \theta, \text{ and } \theta \text{ is a SSOSP of } \mathcal{L}\left(\theta; x, -\frac{\nabla_\theta \mathcal{L}(\theta,x)}{\sigma}\right)\right\} \\
&= \mathbb{1}\left\{\widehat{\theta}\left(x, -\frac{\nabla_\theta \mathcal{L}(\theta;x)}{\sigma}\right) = \theta, \text{ and } \nabla^2_\theta \mathcal{L}(\theta;x) \succ 0\right\}.
\end{aligned}$$

In other words, given the proposed state x, we need only verify (1) that $\nabla^2_\theta \mathcal{L}(\widehat{\theta};x) \succ 0$, which is a simple calculation, and (2) that the estimator $(x,w) \mapsto \widehat{\theta}(x,w)$, when calculated with this proposed x and with $w = -\frac{\nabla_\theta \mathcal{L}(\widehat{\theta};x)}{\sigma}$, indeed returns the observed value $\widehat{\theta}$. We note that, in the special case that $\mathcal{L}$ is strictly convex, then this verification is trivial—if we take the map $(x,w) \mapsto \widehat{\theta}(x,w)$ to be the output of any solver guaranteed to return the unique FOSP (if it exists), then (2) is automatically verified since we know that $\widehat{\theta}$ is a FOSP of $\mathcal{L}(\theta;x,w)$ by definition of w, while (1) holds by strict convexity of $\mathcal{L}$.

D.2.1 Choosing the proposal distribution

To choose the proposal distribution $q_{\widehat{\theta}}(x \,|\, x')$, we will bear in mind the following considerations. First, we need to be able to efficiently draw a sample from $q_{\widehat{\theta}}(\cdot \,|\, x')$. Second, we need to trade off between the following two goals: given our current state X_{curr} and a proposed state $X_{\text{prop}} \sim q_{\widehat{\theta}}(\cdot \,|\, X_{\text{curr}})$,

- The acceptance probability $A_{\widehat{\theta}}(X_{\text{prop}} \,|\, X_{\text{curr}})$ should not be too close to zero.

- There should not be too much similarity or dependence between X_{curr} and X_{prop}.

To illustrate this tradeoff, if we define $q_{\widehat{\theta}}(\cdot \,|\, X_{\text{curr}})$ as the point mass at X_{curr} (i.e., we never move to a new state), then the acceptance probability $A_{\widehat{\theta}}(X_{\text{prop}} \,|\, X_{\text{curr}})$ will be equal to 1 almost surely, but the algorithm will return copies $\widetilde{X}^{(1)} = \cdots = \widetilde{X}^{(M)} = X$, leading to a powerless procedure. On the other hand, if we define $q_{\widehat{\theta}}(\cdot \,|\, X_{\text{curr}})$ to draw X_{prop} to be independent or nearly independent of X_{curr} (for example, $X_{\text{prop}} \sim P_{\widehat{\theta}}$), then it may be hard to ensure that $p_{\widehat{\theta}}(X_{\text{prop}} \,|\, \widehat{\theta})$ is sufficiently large to bound $A_{\widehat{\theta}}(X_{\text{prop}} \,|\, X_{\text{curr}})$ away from zero.

Given the well-known challenge of hyperparameter tuning in the field of MCMC [Roberts and Rosenthal, 2009], we can expect that this will be highly non-trivial and problem-dependent. But one appealing aspect of aCSS testing is that we can tune the MCMC hyperparameters *after* looking at $\widehat{\theta}$ without violating any of our theory. We demonstrate how we did so in our four examples below.

Examples 1, 2, and 4 First, we consider the three examples where our model P_θ for X consists of n independent draws—that is, P_θ is a product distribution with density

$$f_\theta(x) = \prod_{i=1}^{n} f_\theta^{(i)}(x_i).$$

In this setting, we begin by fixing a parameter $s \in \{1, \ldots, n\}$ (we will discuss the choice of s shortly). Then the proposal distribution $q_{\widehat{\theta}}(x|x')$ is defined as follows:

- Draw a subset $\mathcal{S} \subseteq \{1, \ldots, n\}$ of size s, uniformly at random.
- For each $i = 1, \ldots, n$,
  - If $i \in \mathcal{S}$, draw $x_i \sim f_{\widehat{\theta}}^{(i)}(\cdot)$.
  - If $i \notin \mathcal{S}$, set $x_i = x'_i$.

We can see that the parameter s controls the tradeoff—a larger s ensures then the proposed state $x = X_{\text{prop}}$ will not be too similar to the previous state $x' = X_{\text{curr}}$, but a smaller s ensures that the acceptance ratio $A_{\widehat{\theta}}(X_{\text{prop}} \,|\, X_{\text{curr}})$ will not be too low (since, when most entries $i = 1, \ldots, n$ of X_{prop} coincide with those of X_{curr}, the ratio $\frac{p_{\widehat{\theta}}(X_{\text{prop}} \,|\, \widehat{\theta})}{p_{\widehat{\theta}}(X_{\text{curr}} \,|\, \widehat{\theta})}$ should be close to 1).

Next, how can we choose s to balance between these two considerations? For these examples, we will choose s from the data itself. First, we observe that allowing s to depend on $\widehat{\theta}$ does not violate the validity of our procedure. This is because the mechanism $\widetilde{P}_M(\cdot \,|\, X, \widehat{\theta})$ for sampling the copies is only required to satisfy assumption (2.9); it is allowed to depend arbitrarily on $\widehat{\theta}$, as long as exchangeability between X and $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ is not violated. (In particular, this means that we *cannot* use the data X itself to choose s.) We will choose s by simulating the procedure with $\widehat{\theta}$ in place of θ_0:

- Let $\theta_0^{\text{sim}} = \widehat{\theta}$.
- Draw $X^{\text{sim}} \sim P_{\theta_0^{\text{sim}}}$.

- For each candidate choice of s, run Metropolis–Hastings initialized at X^{sim}, and compute the average acceptance probability.
- Repeat for many draws of X^{sim} to get an average acceptance probability $\bar{A}_s$ for each s, and among all values of s such that $\bar{A}_s \geq 0.2$, choose the value of s that maximizes $s\bar{A}_s$ (thus maximizing the expected number of elements that change at each MH step).

With this choice of s, we have completed our $\widehat{\theta}$-dependent definition of the proposal distribution $q_{\widehat{\theta}}(x \,|\, x')$ for this setting. Then we choose L to be at least $\frac{n}{s\bar{A}_s}$ to ensure that (most) entries will be resampled within L steps; in our simulations we chose L to be $\min\{500, \frac{2n}{s\bar{A}_s}\}$ (rounded to an integer).

Example 3 Next we consider the Gaussian spatial process. Here we will again define a parametrized proposal distribution, and will then choose the parameter by simulation. For any $\rho \in (0, 1)$, define the proposal distribution $q_{\widehat{\theta}}(x|x')$ as follows:

- Draw $x_{\text{tmp}} \sim \mathcal{N}(0, \Sigma_{\widehat{\theta}})$.
- Set $x = \rho \cdot x' + \sqrt{1 - \rho^2} \cdot x_{\text{tmp}}$.

As for the examples above, the value of ρ governs the tradeoff—in this case, a smaller ρ ensures then the proposed state $x = X_{\text{prop}}$ will not be too similar to the previous state $x' = X_{\text{curr}}$, but a larger ρ ensures that the acceptance ratio $A_{\widehat{\theta}}(X_{\text{prop}} \,|\, X_{\text{curr}})$ will not be too low. In each trial, we will choose ρ with a simulation, analogous to the choice of s for the other examples:

- Let $\theta_0^{\text{sim}} = \widehat{\theta}$.
- Draw $X^{\text{sim}} \sim P_{\theta_0^{\text{sim}}}$, $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, and calculate $\widehat{\theta} = \widehat{\theta}(X^{\text{sim}}, W)$.
- For each candidate choice of ρ, run one step of Metropolis–Hastings initialized at X^{sim}, to generate X^{new}.
- Repeat for 500 draws of X^{sim} (discarding any draws for which $\widehat{\theta}(X^{\text{sim}}, W)$ is not a SSOSP). Among all values of ρ that achieve average acceptance probability ≥ 0.05, find the value of ρ that minimizes the average correlation between X^{sim} and X^{new}.

With this choice of ρ, writing $\hat{\rho}$ to denote the average correlation between X^{sim} and X^{new}, we then set $L = \min\{500, \frac{20}{1-(\hat{\rho})_+}\}$ (rounded to an integer).

E Details for Figure 1

In this section we give details for the simulation that generated Figure 1, comparing the parametric bootstrap versus co-sufficient sampling for a Gaussian linear model setting as described in Section 1. Recall that the null hypothesis for this example is the model

$$X = \theta \cdot Z + \mathcal{N}(0, \mathbf{I}_n)$$

for some $\theta \in \mathbb{R}$, where $Z \in \mathbb{R}^n$ is a fixed covariate vector. We are interested in testing the alternative hypothesis that X is in fact more strongly associated with some other covariate $Y \in \mathbb{R}^n$, and so our test statistic is given by

$$T(X) = \frac{(X^\top Y)^2}{(X^\top Z)^2}.$$

To generate the data, we choose sample size $n = 100$, and then independently for each $i = 1, \ldots, n$, we generate the triple (X_i, Y_i, Z_i) by taking

$$(Y_i, Z_i) \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

with correlation parameter $\rho = 0.97$, and define

$$X_i = \theta_0 \cdot Z_i + \mathcal{N}(0, 1),$$

where the true parameter is chosen as $\theta_0 = 0$.

Next we run parametric bootstrap and CSS to generate copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ of the data $X \in \mathbb{R}^n$, for $M = 500$. For both methods, the MLE is given by $\widehat{\theta} = (Z^\top Z)^{-1} Z^\top X$. To run the parametric bootstrap, we generate the copies from the distribution with parameter $\theta = \widehat{\theta}$, that is, we define the copies as

$$\widetilde{X}^{(m)}_{\text{boot}} = \widehat{\theta} \cdot Z + V_m,$$

where $V_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$. To run CSS, we instead condition on the MLE $\widehat{\theta}$, and the copies can therefore be generated as

$$\widetilde{X}^{(m)}_{\text{CSS}} = \widehat{\theta} \cdot Z + \text{Proj}_Z^{\perp} \cdot V_m,$$

where again $V_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$.

Finally, we repeat the simulation for 10,000 independent trials to generate the histograms of p-values for each method, as shown in Figure 1.

Acknowledgements

The authors would like to thank Michael Bian for help with some of the computation. The first author was supported by the National Science Foundation via grant DMS–1654076, and by the Office of Naval Research via grant N00014-20-1-2337.

References

Alan Agresti. A survey of exact inference for contingency tables. *Statist. Sci.*, 7(1):131–153, 02 1992. doi: 10.1214/ss/1177011454. URL https://doi.org/10.1214/ss/1177011454.

Alan Agresti. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine*, 20(17-18):2709–2722, 2001.

François Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35, 2014.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Rina Foygel Barber and Mladen Kolar. ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422–3450, 2018.

Maurice Stevenson Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937.

CB Bell. Inference for goodness-of-fit problems with nuisance parameters: (applications to signal detection). *Journal of statistical planning and inference*, 9(3):273–284, 1984.

JI Beltrán-Beltrán and FJ O'Reilly. On goodness of fit tests for the Poisson, negative binomial and binomial distributions. *Statistical Papers*, 60(1):1–18, 2019.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76 (4):633–642, 1989.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Michel Broniatowski and Virgile Caron. Conditional inference in parametric models. *arXiv preprint arXiv:1202.0944*, 2012.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

Alberto Contreras-Cristán, Richard A Lockhart, Michael A Stephens, and Shaun Z Sun. On the use of priors in goodness-of-fit tests. *Canadian Journal of Statistics*, 47(4):560–579, 2019.

David Roxbee Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987.

Persi Diaconis, Susan Holmes, Mehrdad Shahshahani, et al. Sampling from a manifold. In *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, pages 102–125. Institute of Mathematical Statistics, 2013.

James Durbin. Some methods of constructing exact tests. *Biometrika*, 48(1-2):41–65, 1961.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Steinar Engen and Magnar Lillegård. Stochastic simulations conditioned on sufficient statistics. *Biometrika*, 84(1):235–240, 1997.

Arnab Hazra. An exact Kolmogorov–Smirnov test for the negative Binomial distribution with unknown probability of success. *Research & Reviews: Journal of Statistics*, 2(1): 1–13, 2013.

Dongming Huang and Lucas Janson. Relaxing the assumptions of knockoffs by conditioning. *Annals of Statistics*, 2020+. To Appear.

John D Kalbfleisch and David A Sprott. Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society: Series B (methodological)*, 32(2):175–194, 1970.

John E Kolassa. Algorithms for approximate conditional inference. *Statistics and Computing*, 13(2):121–126, 2003.

Athanasios Kousathanas, Christoph Leuenberger, Jonas Helfer, Mathieu Quinodoz, Matthieu Foll, and Daniel Wegmann. Likelihood-free inference in high-dimensional models. *Genetics*, 203(2):893–904, 2016.

A Kumar and PK Pathak. Sufficiency and tests of goodness of fit. *Scandinavian Journal of Statistics*, pages 39–43, 1977.

Kenneth L Lange, Roderick JA Little, and Jeremy MG Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Lucien Le Cam. Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.*, 3:37–98, 1960.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475035.

Magnar Lillegård. Tests based on Monte Carlo simulations conditioned on maximum likelihood estimates of nuisance parameters. *Journal of statistical computation and simulation*, 71(1):1–10, 2001.

Magnar Lillegård and Steinar Engen. Exact confidence intervals generated by conditional parametric bootstrapping. *Journal of Applied Statistics*, 26(4):447–459, 1999.

Bo H Lindqvist and Bjarte Rannestad. Monte Carlo exact goodness-of-fit tests for nonhomogeneous Poisson processes. *Applied Stochastic Models in Business and Industry*, 27(3): 329–341, 2011.

Bo Henry Lindqvist and Gunnar Taraldsen. Exact statistical inference for some parametric nonhomogeneous Poisson processes. *Journal of The Iranian Statistical Society*, 12(1): 113–126, 2013.

Richard A Lockhart. Conditional limit laws for goodness-of-fit tests. *Bernoulli*, 18(3):857–882, 2012.

Richard A Lockhart, Federico J O'Reilly, and Michael A Stephens. Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika*, 94(4):992–998, 2007.

Richard A Lockhart, Federico O'Reilly, and Michael Stephens. Exact conditional tests and approximate bootstrap tests for the von Mises distribution. *Journal of Statistical Theory and Practice*, 3(3):543–554, 2009.

Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Federico O'Reilly and Leticia Gracia-Medrano. On the conditional distribution of goodness-of-fit tests. *Communications in Statistics-Theory and Methods*, 35(3):541–549, 2006.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

James D Santos and Nelson L Souza Filho. A Metropolis algorithm to obtain co-sufficient samples with applications in conditional tests. *Communications in Statistics-Simulation and Computation*, 48(9):2655–2659, 2019.

Michael A Stephens. Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14(3):785–791, 2012.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *Ann. Statist.*, 46(2):679–710, 04 2018. doi: 10.1214/17-AOS1564. URL https://doi.org/10.1214/17-AOS1564.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.