
A Black-Box Debiasing Framework for Conditional Sampling

Han Cui

University of Illinois at Urbana-Champaign
Champaign, IL
hancui5@illinois.edu

Jingbo Liu

University of Illinois at Urbana-Champaign
Champaign, IL
jingbo1@illinois.edu

Abstract

Conditional sampling is a fundamental task in Bayesian statistics and generative modeling. Consider the problem of sampling from the posterior distribution $P_{X|Y=y^*}$ for some observation y^* , where the likelihood $P_{Y|X}$ is known, and we are given n i.i.d. samples $D = \{X_i\}_{i=1}^n$ drawn from an unknown prior distribution π_X . Suppose that $f(\hat{\pi}_{X^n})$ is the distribution of a posterior sample generated by an algorithm (e.g. a conditional generative model or the Bayes rule) when $\hat{\pi}_{X^n}$ is the empirical distribution of the training data. Although averaging over the randomness of the training data D , we have $\mathbb{E}_D(\hat{\pi}_{X^n}) = \pi_X$, we do not have $\mathbb{E}_D\{f(\hat{\pi}_{X^n})\} = f(\pi_X)$ due to the nonlinearity of f , leading to a bias. In this paper we propose a black-box debiasing scheme that improves the accuracy of such a naive plug-in approach. For any integer k and under boundedness of the likelihood and smoothness of f , we generate samples $\hat{X}^{(1)}, \dots, \hat{X}^{(k)}$ and weights $w_1, \dots, w_k$ such that $\sum_{i=1}^k w_i P_{\hat{X}^{(i)}}$ is a k -th order approximation of $f(\pi_X)$, where the generation process treats f as a black-box. Our generation process achieves higher accuracy when averaged over the randomness of the training data, without degrading the variance, which can be interpreted as improving memorization without compromising generalization in generative models.

1 Introduction

Conditional sampling is a major task in Bayesian statistics and generative modeling. Given an observation y^* , the objective is to draw samples from the posterior distribution $P_{X|Y=y^*}$, where the likelihood $P_{Y|X}$ is known but the prior distribution π_X is unknown. Instead, we are provided with a dataset $D = \{X_i\}_{i=1}^n$ consisting of n i.i.d. samples drawn from π_X .

The setting is common in a wide range of applications, including inpainting and image deblurring [9, 5] (where X is an image and $Y|X$ is a noisy linear transform), text-conditioned image generation [7, 13] (where X is an image and Y is a natural language prompt), simulating biomedical structures with desired properties, and trajectory simulations for self-driving cars. Moreover, conditional sampling is equally vital in high-impact machine learning and Bayesian statistical methods, particularly under distribution shift, such as in transfer learning. For instance, conditional sampling has enabled diffusion models to generate trajectories under updated policies, achieving state-of-the-art performance in offline reinforcement learning [8, 1, 26]. Pseudo-labeling, a key technique for unsupervised pretraining [10] and transfer learning calibration [20], relies on generating conditional labels. Additionally, conditional diffusion models seamlessly integrate with likelihood-free inference [6, 18, 27]. Existing approaches often use generative models such as *VAEs* or *Diffusion models* to generate samples by learning $P_{X|Y=y^*}$ implicitly from the data.

Our work focuses on approximating the true posterior $P_{X|Y=y^*}$ using the observed samples $D = X^n = (X_1, \dots, X_n)$ and the new observation y^* , but without the knowledge of the prior. Denote by $P_{\hat{X}|Y=y^*, D}$ the approximating distribution. We can distinguish two kinds of approximations: First, $P_{\hat{X}|Y=y^*, D} \approx P_{X|Y=y^*}$ with high probability over D , which captures the *generalization* ability since the model must learn the distribution from the training samples. This criterion is commonly adopted in estimation theory and has also been examined in the convergence analysis of generative models [16, 28, 26, 22]. Second, $\mathbb{E}_D(P_{\hat{X}|Y=y^*, D}) \approx P_{X|Y=y^*}$ is a weaker condition since it only requires approximation when averaged over the randomness of the training data, but is still useful in some sampling and generative tasks, e.g. generating samples for bootstrapping or Monte Carlo estimates of function expectations. The second condition captures the ability to *memorize* or *imitate* training sample distribution. It is interesting to note that in the unconditional setting (i.e., without distribution shift), a permutation sampler can perfectly imitate the unknown training data distribution, even if $n = 1$, so the problem is trivial from the sample complexity perspective. However, in the conditional setting, it is impossible to get such a perfect imitation with finite training data, as a simple binary distribution example in Section 3.2 illustrates. It naturally leads to the following question:

How fast can the posterior approximation converge to the true posterior as $n \rightarrow \infty$, and is there a sampling scheme achieving this convergence rate?

Contribution. We address the question above by proposing a novel debiasing framework for posterior approximation. Our main contributions can be summarized as follows:

- **Debiasing framework for posterior approximation.** We introduce a novel debiasing framework for posterior approximation with an unknown prior. Our method leverages the known likelihood $P_{Y|X}$ and the observed data to construct an improved approximate posterior $\tilde{P}_{X^n}(x|y^*)$ with provably reduced bias. In particular, let $f(\hat{\pi}_{X^n})$ represent the distribution of a posterior sample generated by an algorithm f when $\hat{\pi}_{X^n}$ is the empirical distribution of the training data. Then for any integer k , assuming that the likelihood function $P_{Y|X}$ is bounded and f is sufficiently smooth, we generate samples $\hat{X}^{(1)}, \dots, \hat{X}^{(k)}$ from f based on multiple resampled empirical distributions. These are then combined with designed (possibility negative) weights $w_1, \dots, w_k$ to construct an approximate posterior:

$$\tilde{P}_{X^n}(\cdot|y^*) = \sum_{i=1}^k w_i P_{\hat{X}^{(i)}}$$

which is a k -th order approximation of $f(\pi_X)$, treating the generation process f as a black-box. Our generation process achieves higher accuracy when averaged over the randomness of the training data, but not conditionally on the training data, which highlights the trade-off between memorization and generalization in generative models. Specifically, we do not assume any parametric form for the prior and our method can achieve a bias rate of $\mathcal{O}(n^{-k})$ for any prescribed integer k and a variance rate of $\mathcal{O}(n^{-1})$.

- **Theoretical bias and variance guarantees.** We establish theoretical guarantees on both bias and variance for the Bayes-optimal sampler under continuous prior setting and for a broad class of samplers f with a continuous $2k$ -th derivative, as specified in Assumption 2, under the discrete prior setting. The proposed debiasing framework can also be applied in a black-box manner (see Remark 2 for the intuition), making it applicable to a broad class of state-of-the-art conditional samplers, such as diffusion models and conditional VAE. Based on this perspective, we treat the generative model f as a black box that can output posterior samples given resampled empirical distributions. Applying f to multiple recursive resampled versions of the training data and combining the outputs with polynomial weights, we obtain a bias-corrected approximation of the posterior. The procedure is described in Algorithm 1.

Our approach is also related to importance sampling. Since the true posterior $P_{X|Y}$ is intractable to compute, we can use expectations under the debiased posterior $\tilde{P}_{X^n}(x|y^*)$ to approximate the expectations under the true posterior $P_{X|Y=y^*}$. For a test function h , we estimate $\mathbb{E}_{P_{X|Y=y^*}}\{h(X)\}$ by

$$\mathbb{E}_{\tilde{P}_{X^n}(x|y^*)}\{h(X)\} \approx \frac{1}{N} \sum_{i=1}^N h(\tilde{X}_j) \frac{\tilde{P}_{X^n}(\tilde{X}_j|y^*)}{q(\tilde{X}_j|y^*)}, \quad (1)$$

Algorithm 1 Posterior Approximation via Debiasing Framework

Input: Observation y^* , likelihood $P_{Y|X}$, data $X^n = (X_1, \dots, X_n)$, number of steps k , a black-box conditional sampler f (i.e., a map from a prior distribution to a posterior distribution)

Output: $\hat{X}^{(j)}, j = 1, \dots, k$ such that $\sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j P_{\hat{X}^{(j+1)}}$ is a high-order approximation of the posterior $P_{X|Y=y^*}$

- 1: Initialize $\hat{p}^{(1)} := \hat{\pi}_{X^n}$
 - 2: **for** $\ell = 2$ to k **do**
 - 3: Generate n i.i.d. samples from $\hat{p}^{(\ell-1)}$
 - 4: Let $\hat{p}^{(\ell)}$ be the empirical distribution of the sampled data
 - 5: **end for**
 - 6: **for** $j = 1$ to k **do**
 - 7: Generate samples $\hat{X}^{(j)} \sim f(\hat{p}^{(j)})$
 - 8: **end for**
 - 9: Return $\hat{X}^{(j)}, j = 1, \dots, k$
-

where $\tilde{X}_j \sim q(x|y^*)$ for a chosen proposal distribution q . This resembles our method, in which we approximate the true posterior by a weighted combination $\sum_{i=1}^k w_i P_{\tilde{X}^{(i)}}$. And in (1), the term $\tilde{P}_{X^n}(\tilde{X}_j|y^*)/q(\tilde{X}_j|y^*)$ can be interpreted as a weight assigned to each sample, analogous to the weights w_i in our framework. Therefore, we expect that Algorithm 1 can be broadly applied to Monte Carlo estimates of function expectations, similar to the standard importance sampling technique.

2 Related work

Jackknife Technique. Our work is related to the jackknife technique [17], a classical method for bias reduction in statistical estimation that linearly combines estimators computed on subsampled datasets. Specifically, the jackknife technique generates leave-one-out (or more generally, leave- s -out where $s \geq 1$) versions of an estimator, and then forms a weighted combination to cancel lower-order bias terms. Recently, Nowozin [14] applied the jackknife to the importance-weighted autoencoder (IWAE) bound $\hat{\mathcal{L}}_n$, which estimates the marginal likelihood $\log \pi(x)$ using n samples. While $\hat{\mathcal{L}}_n$ is proven to be an estimator with bias of order $\mathcal{O}(n^{-1})$, the jackknife correction produces a new estimator with reduced bias of order $\mathcal{O}(n^{-m})$. Our paper introduces a debiasing framework based on the similar idea that using a linear combination of multiple approximations to approximate the posterior.

Conditional Generative Models. Conditional generative models have become influential and have been extensively studied for their ability to generate samples from the conditional data distribution $P(\cdot|y)$ where y is the conditional information. This framework is widely applied in vision generation tasks such as text-to-image synthesis [13, 24, 2] where y is an input text prompt, and image inpainting [11, 21] where y corresponds to the known part of an image. We expect that our proposed debiasing framework could work for a broad class of conditional generative models to construct a high order approximation of the posterior $P(\cdot|y)$.

Memorization in Generative Models. The trade-off between memorization and generalization has been a focus of research in recent years. In problems where generating new structures or preserving privacy of training data is of high priority, generalization is preferred over memorization. For example, a study by Carlini et al. [4] demonstrates that diffusion models can unintentionally memorize specific images from their training data and reproduce them when generating new samples. To reduce the memorization of the training data, Somepalli et al. [19] applies randomization and augmentation techniques to the training image captions. Additionally, Yoon et al. [25] investigates the connection between generalization and memorization, proposing that these two aspects are mutually exclusive. Their experiments suggest that diffusion models are more likely to generalize when they fail to memorize the training data. On the other hand, memorizing and imitating the training data may be intentionally exploited, if the goal is Monte Carlo sampling for evaluations of expected values, or if the task does not involve privacy issues, e.g. image inpainting and reconstruction. In these applications, the ability to imitate or memorize the empirical distribution of the training data becomes essential, especially when generalization is unattainable due to the insufficient data. Our work focuses

on memorization phase and shows that it is possible to construct posterior approximations with provably reduced bias by exploiting the empirical distribution.

Mixture-Based Approximation of Target Distributions. Sampling from a mixture of distributions $a_1 P_{X_1} + a_2 P_{X_2} + \dots + a_k P_{X_k}$ to approximate a target distribution P^* is commonly used in Bayesian statistics, machine learning, and statistical physics, especially when individual samples or proposals are poor approximations, but their ensemble is accurate. Traditional importance sampling methods often rely on positive weights, but recent work has expanded the landscape to include more flexible and powerful strategies, including the use of signed weights and gradient information. For example, Oates et al. [15] uses importance sampling and control functional estimators to construct a linear combination of estimators with weights a_k to form a variance-reduced estimator for an expectation under a target distribution P^* . Liu and Lee [12] select the weights a_k by minimizing the empirical version of the kernelized Stein discrepancy (KSD), which often results in negative weights.

3 Problem setup and notation

Consider a dataset $\{X_i\}_{i=1}^n$ consisting of n independent and identically distributed (i.i.d.) samples, where $X_i \in \mathcal{X}$ is drawn from an unknown prior distribution π_X and the conditional distribution $P_{Y|X}$ is assumed to be known. In the Bayesian framework, the posterior distribution of X given Y is given by

$$P_{X|Y}(dx|y) = \frac{P_{Y|X}(y|x)\pi_X(dx)}{\int P_{Y|X}(y|x)\pi_X(dx)}.$$

Given the observed data $X^n = (X_1, \dots, X_n)$ and the new observation y^* , our goal is to approximate the true posterior $P_{X|Y=y^*}$.

3.1 Naive plug-in approximation

A natural approach is to replace the unknown prior π_X with its empirical counterpart

$$\hat{\pi}_{X^n} = n^{-1} \sum_{i=1}^n \delta_{X_i}$$

in the Bayes' rule to compute an approximate posterior which yields the plug-in posterior

$$\hat{P}_{X|Y}(dx|y^*) = \frac{P_{Y|X}(y^*|x)\hat{\pi}_{X^n}(dx)}{\int P_{Y|X}(y^*|x)\hat{\pi}_{X^n}(dx)}. \quad (2)$$

Note that even though $\mathbb{E}_D(\hat{\pi}_{X^n}) = \pi_X$, the nonlinearity of Bayes' rule makes the resulting posterior (2) still biased, that is, $\mathbb{E}_D\{\hat{P}_{X|Y}(\cdot|y^*)\} \neq P_{X|Y}(\cdot|y^*)$. If the denominator in (2) were replaced with $\int P_{Y|X}(y^*|x)\pi_X(dx)$, then averaging the R.H.S. of (2) over the randomness in X^n would yield the true posterior $P_{X|Y}(dx|y^*) = P_{Y|X}(y^*|x)\pi_X(dx) / \int P_{Y|X}(y^*|x)\pi_X(dx)$ exactly.

For typical choices of $P_{Y|X}$ which have nice conditional density (e.g., the additive Gaussian noise channel), $\int P_{Y|X}(y^*|x)\hat{\pi}_{X^n}(dx)$ converges at the rate of $n^{-1/2}$, by the central limit theorem. Consequently, $\mathbb{E}_D(\hat{P}_{X|Y=y^*})$ converges to the true posterior at the rate $\tilde{O}(n^{-1/2})$ in the ∞ -Renyi divergence metric regardless of the smoothness of π_X . Under appropriate regularity conditions, we can in fact show that $\mathbb{E}_D(\hat{P}_{X|Y=y^*})$ converges at the rate of $\tilde{O}(n^{-1})$, which comes from the variance term in the Taylor expansion. Naturally, we come to an essential question: can we eliminate the bias entirely? That is, is it possible that $\mathbb{E}_D\{\hat{P}_{X|Y}(\cdot|y^*)\} = P_{X|Y}(\cdot|y^*)$?

3.2 Impossibility of exact unbiasedness

Exact unbiasedness is, in general, unattainable. Consider the simple case where X is binary, that is, $X \sim \text{Bern}(q)$ for some unknown parameter $q \in (0, 1)$. Define the likelihood ratio $\alpha = \alpha(y^*) :=$

$P_{Y|X}(y^*|1)/P_{Y|X}(y^*|0)$. Then the true posterior is

$$X|Y = y^* \sim \text{Bern}\left(\frac{\alpha q}{\alpha q + 1 - q}\right).$$

On the other hand, if we approximate the posterior distribution as $\hat{P}_{X|Y}(1|y = y^*) = \text{Bern}(p(k))$ upon seeing k outcomes equal to 1, then

$$\mathbb{E}_D \left\{ \hat{P}_{X|Y}(1|y^*) \right\} = \sum_{k=0}^n p(k) \binom{n}{k} q^k (1-q)^{n-k}, \quad (3)$$

which is a polynomial function of q , and hence cannot equal the rational function $\alpha q/(\alpha q + 1 - q)$ for all q . This implies that an *exact* imitation, in the sense that $\mathbb{E}_D \{ \hat{P}_{X|Y}(\cdot|y^*) \} = P_{X|Y}(\cdot|y^*)$, $\forall \pi_X$, is impossible. However, since a rational function can be *approximated* arbitrarily well by polynomials, this does not rule out the possibility that a better sampler achieving convergence faster than, say, the $\tilde{\mathcal{O}}(n^{-1/2})$ rate of the naive plug-in method. Indeed, in this paper we propose a black-box method that can achieve convergence rates as fast as $\mathcal{O}(n^{-k})$ for any fixed $k > 0$.

3.3 Objective and notation

Since the bias in the plug-in approximation arises from the nonlinearity of Bayes' rule, we aim to investigate whether a faster convergence rate can be achieved. Our objective is to construct an approximation $\tilde{P}_{X^n}(x|y = y^*)$ that improves the plug-in approximation by reducing the bias. Specifically, the debiased approximation satisfies the following condition:

$$\left| \mathbb{E}_{X^n} \left\{ \tilde{P}_{X^n}(x|y = y^*) \right\} - P_{X|Y}(x|y^*) \right| < \left| \mathbb{E}_{X^n} \left\{ \hat{P}_{X|Y}(x|y^*) \right\} - P_{X|Y}(x|y^*) \right|.$$

More generally, we can replace the Bayes rule by an arbitrary map f from a prior to a posterior distribution (e.g. by a generative model), and the goal is to construct a debiased map $\tilde{f}$ such that

$$\left\| \mathbb{E}_{X^n} \tilde{f}(\hat{\pi}_{X^n}) - f(\pi) \right\|_{\text{TV}} < \left\| \mathbb{E}_{X^n} f(\hat{\pi}_{X^n}) - f(\pi) \right\|_{\text{TV}}.$$

Notation. Let δ_x denote the Dirac measure, $\|\cdot\|_{\text{TV}}$ denote the total variation norm. For any positive integer m , denote $[m] = \{1, \dots, m\}$ as the set of all positive integers smaller than all equal to m . Write $b_n = \mathcal{O}(a_n)$ if b_n/a_n is bounded as $n \rightarrow \infty$. Write $b_n = \mathcal{O}_s(a_n)$ if b_n/a_n is bounded by $C(s)$ as $n \rightarrow \infty$ for some constant $C(s)$ that depends only on s . We use the notation $a \lesssim b$ to indicate that there exists a constant $C > 0$ such that $a \leq Cb$. Similarly, $a \lesssim_k b$ means that there exists a constant $C(k) > 0$ that depends only on k such that $a \leq C(k)b$. Furthermore, for notational simplicity, we will use π to denote the true prior π_X and $\hat{\pi}$ to denote the empirical prior $\hat{\pi}_{X^n}$ in the rest of the paper.

4 Main result

4.1 Debiased posterior approximation under continuous prior

Let $\Delta_{\mathcal{X}}$ denote the space of probability measures on $\mathcal{X}$. Define the likelihood function $\ell(x) := P_{Y|X}(y^*|x)$, which represents the probability of observing the data y^* given x . Let $f : \Delta_{\mathcal{X}} \rightarrow \Delta_{\mathcal{X}}$ be a map from the prior measure to the posterior measure, conditioned on the observed data y^* . Let B_n be the operator such that for any function $f : \Delta_{\mathcal{X}} \rightarrow \Delta_{\mathcal{X}}$,

$$B_n f(p) = \mathbb{E} \{ f(\hat{p}) \}, \quad (4)$$

where $\hat{p}$ denotes the empirical measure of n i.i.d. samples from measure p .

We consider the case that f represents a mapping corresponding to the Bayes posterior distribution. Using Bayes' theorem, for any measure $\pi \in \Delta_{\mathcal{X}}$ and any measurable set $A \subset \mathcal{X}$, the posterior measure $f(\pi)$ is expressed as

$$f(\pi)(A) = \frac{\int_A \ell(x) \pi(dx)}{\int_{\mathcal{X}} \ell(x) \pi(dx)}.$$

As discussed in Section 3, the equality $B_n f(\pi) = f(\pi)$ is not possible due to the nonlinearity of f . However, we can achieve substantial improvements over the plug-in method by using polynomial approximation techniques analogous to those from prior statistical work by Cai and Low [3] and Wu and Yang [23]. For $k \geq 1$, we define the operator $D_{n,k}$ as a linear combination of the iterated operators B_n^j for $j = 0, \dots, k-1$:

$$D_{n,k} = \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j B_n^j.$$

Assumption 1. *The likelihood function ℓ is bounded, i.e., there exists $0 < L_1 \leq L_2$ such that $L_1 \leq \ell(x) \leq L_2$.*

The following theorem provides a systematic method for constructing an approximation of $f(\pi)$ with an approximation error of order $\mathcal{O}(n^{-k})$ for any desired integer k .

Theorem 1. *Under Assumption 1, for any measurable set $A \subset \mathcal{X}$ and any $k \in \mathbb{N}^+$, we have*

$$\|\mathbb{E}_{X^n} \{D_{n,k} f(\hat{\pi})\} - f(\pi)\|_{\text{TV}} = \mathcal{O}_{L_1, L_2, k}(n^{-k}), \quad (5)$$

$$\text{Var}_{X^n} \{D_{n,k} f(\hat{\pi})(A)\} = \mathcal{O}_{L_1, L_2, k}(n^{-1}). \quad (6)$$

Remark 1. $D_{n,k} f(\hat{\pi}) = \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j B_n^j f(\hat{\pi})$ in (5) can be interpreted as a weighted average of the distribution of some samples. Specifically, if we treat the coefficient $\binom{k}{j+1} (-1)^j$ as the weight w_j and $B_n^j f(\hat{\pi})$ as the distribution of some sample $\hat{X}^{(j)}$, then $D_{n,k} f(\hat{\pi}) = \sum_{j=0}^{k-1} w_j P_{\hat{X}^{(j)}}$.

Remark 2. Recall the binary case discussed in Section 3, (3) illustrates that we cannot get an exact approximation for the true posterior. But from (5), we demonstrate that even if $\|\mathbb{E}_{X^n} \{D_{n,k} f(\hat{\pi})\} - f(\pi)\|_{\text{TV}} = 0$ is impossible, it can be arbitrarily small. Although the theoretical guarantees are derived for the Bayes-optimal sampler, (5) is expected to hold for general sampler f such as diffusion models. Here we give the intuition for this conjecture. We view the operator $B_n f(\pi) := \mathbb{E}\{f(\hat{\pi})\}$ as a good approximation of $f(\pi)$, i.e., $B_n \approx I$, where I is the identity operator. This implies that the error operator $E := I - B_n$ is a “small” operator. Under this heuristic, if $E f(\pi) = \mathcal{O}(n^{-1})$, intuitively we have $E^k f(\pi) = \mathcal{O}(n^{-k})$. Using the binomial expansion of $E^k = (I - B_n)^k$, we have $E^k f(\pi) = f(\pi) - \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^j f(\pi) = f(\pi) - \mathbb{E}\{\sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^{j-1} f(\hat{\pi})\} = f(\pi) - \mathbb{E}\{D_{n,k} f(\hat{\pi})\}$. This representation motivates the specific form of $D_{n,k}$.

Remark 3. In general, the curse of dimensionality may arise and depends on the specific distribution of X and the likelihood function ℓ . There is no universal relationship between n and the dimension d . However, to build intuition, we give an example that illustrates how n and d may relate. Suppose that $Y = (Y(1), \dots, Y(d))$ and $X = (X(1), \dots, X(d))$ have i.i.d. components, and $L_1 \leq P(Y(i)|X(i)) \leq L_2$ for $1 \leq i \leq d$. Then we have $\ell(X) := P(Y|X) \in [L_1^d, L_2^d]$. Note that $\mathcal{O}_{L_1, L_2, k}(n^{-k})$ in (5) can be bounded by $C(k)(L_2^d/L_1^d)^{2k} n^{-k}$ for some constant $C(k)$ depending only on k . To ensure that our debiasing method improves over the baseline method without debiasing in the case of growing dimensions, it suffices to let n and d satisfy that $(L_2^d/L_1^d)^{2k} n^{-k} \ll n^{-1}$ when $k \geq 2$, which is equivalent to $kd \ll \log(n)$.

Sketch proof for Theorem 1. First let $\mu = \int_{\mathcal{X}} \ell(x) \pi(dx)$ and $\mu_A = \int_A \ell(x) \pi(dx)$ and introduce a new operator

$$C_{n,k} := \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^j,$$

then we have $B_n D_{n,k} = C_{n,k}$. By the definition of B_n , it suffices to show that

$$C_{n,k} f(\pi)(A) - f(\pi)(A) = \mathcal{O}_{L_1, L_2, k}(n^{-k}).$$

The first step is to express $B_n^j f(\pi)$ with the recursive resampled versions of the training data. Specifically, let $\hat{\pi}^{(0)} = \pi$, $\hat{\pi}^{(1)} = \hat{\pi}$ and set $(X_1^{(0)}, \dots, X_n^{(0)}) \equiv (X_1, \dots, X_n)$. For $j = 1, \dots, k$,

we define $\hat{\pi}^{(j)}$ as the empirical measure of n i.i.d. samples $(X_1^{(j-1)}, \dots, X_n^{(j-1)})$ drawn from the measure $\hat{\pi}^{(j-1)}$. Additionally, let

$$e_n^{(j)} = n^{-1} \sum_{i=1}^n \{\ell(X_i^{(j)}) - \mu\} \quad \text{and} \quad \mu_A^{(j)} = n^{-1} \sum_{i=1}^n \ell(X_i^{(j)}) \delta_{X_i^{(j)}}(A).$$

Then we have

$$C_{n,k}f(\pi)(A) = \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^j f(\pi)(A) = \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left(\frac{\mu_A^{(j-1)}}{e_n^{(j-1)} + \mu} \right). \quad (7)$$

The second step is to rewrite (7) with Taylor expansion of $\mu_A^{(j-1)} / (e_n^{(j-1)} + \mu)$ with respect to $e_n^{(j-1)}$ up to order $2k - 1$. $L_1 \leq l(X_i^{(j-1)}) \leq L_2$ and Hoeffding's inequality implies that the expectation of the residual term $\mathbb{E}\{(e_n^{(j-1)})^{2k} / \xi^{2k+1}\}$ for some ξ between $e_n^{(j-1)} + \mu$ and μ is $\mathcal{O}_{L_1, L_2, k}(n^{-k})$. Now we instead to show that

$$B_{k,r} := \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left\{ \mu_A^{(j-1)} (e_n^{(j-1)})^r \right\} = \mathcal{O}_{L_1, L_2, k}(n^{-k}),$$

since (7) is equal to $\mu_A / \mu + \sum_{r=1}^{2k-1} (-1)^r \mu^{-r-1} B_{k,r} + \mathcal{O}_{L_1, L_2, k}(n^{-k})$.

Define a new operator $B : h \mapsto \mathbb{E}[h(\hat{\pi})]$ for any $h : \Delta_{\mathcal{X}} \rightarrow \mathbb{R}$ and let $h_s(\pi) = \{\int_A \ell(x) \pi(dx)\} \{\int \ell(x) \pi(dx)\}^s$. Then

$$B_{k,r} = \sum_{s=0}^r \binom{r}{s} (-1)^{(r-s)} \mu^{r-s} \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B^j h_s(\pi).$$

The last step is to prove

$$(I - B)^k h_s(\pi) = \mathcal{O}_{L_1, L_2, s}(n^{-k}), \quad (8)$$

since (8) is equivalent to $\sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B^j h_s(\pi) = h_s(\pi) + \mathcal{O}_{L_1, L_2, s}(n^{-k})$. Finally (8) follows from the fact that $(I - B)^k h_s(\pi)$ can be expressed as a finite sum of the terms which have the following form:

$$\alpha_{\mathbf{a}, \mathbf{s}, v} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i},$$

where $|\alpha_{\mathbf{a}, \mathbf{s}, v}| \leq C_k(s) n^{-k}$ for some constant $C_k(s)$ (see our Lemma 2).

□

4.2 Debiased posterior approximation under discrete prior

In this section, we consider the case where X follows a discrete distribution. As mentioned in Remark 2, the result in Theorem 1 is expected to hold in a broader class of samplers f under smoothness, extending beyond just the Bayes-optimal sampler f . The assumption of finite $\mathcal{X}$ in this section allows us to simplify some technical aspects in the proof.

Let the support of X be denoted as $\mathcal{X} = \{u_1, u_2, \dots, u_m\}$. Assume that $|\mathcal{X}| = m$ is finite, and X is distributed according to an unknown prior distribution $\pi(x)$ such that the probability of X taking the value u_i is given by $\pi(X = u_i) = q_i$ for $i = 1, 2, \dots, m$. Here, the probabilities q_i are unknown and satisfy the usual constraints that $q_i \geq 0$ for all i and $\sum_{i=1}^m q_i = 1$.

Let $\mathbf{q} = (q_1, \dots, q_m)^\top$ represent the true prior probability vector associated with the probability distribution $\pi(x)$. Let $\mathbf{g}$ be a map from a prior probability vector to a posterior probability vector. Then $\mathbf{g}(\mathbf{q}) = (g_1(\mathbf{q}), \dots, g_m(\mathbf{q}))^\top$ is the probability vector associated with the posterior. Let $\mathbf{T} = (T_1, \dots, T_m)^\top$ where $T_j = \sum_{i=1}^n \mathbb{1}_{X_i = u_j}$ for $j = 1, \dots, m$. In such setting, by the definition (4) of operator B_n , we can rewrite the operator B_n as

$$B_n g_s(\mathbf{q}) = \mathbb{E} \{g_s(\mathbf{T}/n)\} = \sum_{\nu \in \bar{\Delta}_m} g_s\left(\frac{\nu}{n}\right) \binom{n}{\nu} \mathbf{q}^\nu,$$

where $\bar{\Delta}_m = \{\boldsymbol{\nu} \in \mathbb{N}^m : \sum_{j=1}^m \nu_j = n\}$ and

$$\binom{n}{\boldsymbol{\nu}} = \frac{n!}{\nu_1! \cdots \nu_m!}, \quad \mathbf{q}^\nu = q_1^{\nu_1} \cdots q_m^{\nu_m}.$$

Additionally, let $\Delta_m = \{\mathbf{q} \in \mathbb{R}^m : q_j \geq 0, \sum_{j=1}^m q_j = 1\}$ and let $\|\cdot\|_{C^k(\Delta_m)}$ denote the $C^k(\Delta_m)$ -norm which is defined as $\|f\|_{C^k(\Delta_m)} = \sum_{\|\alpha\|_1 \leq k} \|\partial^\alpha f\|_\infty$ for any $f \in C^k(\Delta_m)$.

Assumption 2. $|\mathcal{X}| = m$ is finite, and $\max_{s \in [m]} \|g_s\|_{C^{2k}(\Delta_m)} \leq G$ for some constant G .

The following theorem provides a systematic method for constructing an approximation of $g_s(\mathbf{q})$ with an error of order $\mathcal{O}(n^{-k})$ for any desired integer k .

Theorem 2. If $|\mathcal{X}| = m$, let $\mathbf{q} = (q_1, \dots, q_m)^\top$ be the true prior probability vector associated with a discrete probability distribution and $\mathbf{T} = (T_1, \dots, T_m)^\top$ where $T_j = \sum_{i=1}^n \mathbb{1}_{X_i=u_j}$ for $j = 1, \dots, m$. Under Assumption 2, the following holds for any $s \in \{1, \dots, m\}$ and any $k \in \mathbb{N}^+$:

$$\begin{aligned} \mathbb{E}_{X^n} \{D_{n,k}(g_s)(\mathbf{T}/n)\} - g_s(\mathbf{q}) &= \mathcal{O}_{k,m,G}(n^{-k}), \\ \text{Var}_{X^n} \{D_{n,k}(g_s)(\mathbf{T}/n)\} &= \mathcal{O}_{k,m,G}(n^{-1}). \end{aligned}$$

Theorem 2 follows directly from the following lemma, which provides the key approximation result.

Lemma 1. For any integers $k, m \in \mathbb{N}^+$ and any function $f \in C^k(\Delta_m)$, we have

$$\|C_{n, \lceil k/2 \rceil}(f) - f\|_\infty = \|(B_n - I)^{\lceil k/2 \rceil}(f)\|_\infty \lesssim_{k,m} \|f\|_{C^k(\Delta_m)} n^{-k/2}.$$

Note that Theorem 2 holds for all mappings $\mathbf{g}$ that satisfy Assumption 2. When $\mathbf{g}$ represents a mapping corresponding to the Bayes posterior distribution, we know the exact form of $\mathbf{g}(\mathbf{q})$. Hence, we can explore sampling schemes for Bayes-optimal mapping $\mathbf{g}$.

We claim that Bayes-optimal mapping $\mathbf{g}$ satisfies Assumption 2. In fact, let $\ell_s = \ell(u_s) := P_{Y|X}(y^*|u_s)$. Using Bayes' theorem, the posterior probability of $X = u_s$ given y^* is given by

$$P_{X|Y}(u_s|y^*) = \frac{\ell_s q_s}{\sum_{j=1}^m \ell_j q_j}.$$

In this case, $g_s(\mathbf{q}) := \ell_s q_s / \sum_{j=1}^m \ell_j q_j$ for $s = 1, \dots, m$. Since $|\mathcal{X}| = m$ is finite, we know that there exists a constant $c_1, c_2 > 0$ such that $c_1 \leq \ell_j \leq c_2$ for all $1 \leq j \leq m$, which implies that $\max_{s \in [m]} \|g_s\|_{C^{2k}(\Delta_m)} \leq G$ for some constant G based on k .

Moreover, estimating $g_s(\mathbf{q})$ based on the observations of $X^n = (X_1, \dots, X_n)$ and y^* is sufficient to generate samples from the posterior distribution $P_{X|Y}(u_s|y^*)$ for $s = 1, \dots, m$. Since the exact form of g_s is known, if we let $\tilde{P}_{X^n}(x = u_s|y = y^*) = D_{n,k}(g_s)(\mathbf{T}/n)$ where $\mathbf{T}/n$ denotes the empirical of the training set, we obtain the following theorem.

Theorem 3. Under Assumption 2, for any $k \in \mathbb{N}^+$, if $|\mathcal{X}| = m$ is finite, then there exists an approximate posterior $\tilde{P}_{X^n}(x|y = y^*)$ satisfies that for any $s \in \{1, \dots, m\}$,

$$\begin{aligned} \mathbb{E}_{X^n} \left\{ \tilde{P}_{X^n}(x = u_s|y = y^*) \right\} - P_{X|Y}(u_s|y^*) &= \mathcal{O}_{k,m,G}(n^{-k}), \\ \text{Var}_{X^n} \left\{ \tilde{P}_{X^n}(x = u_s|y = y^*) \right\} &= \mathcal{O}_{k,m,G}(n^{-1}). \end{aligned}$$

The proposed sampling scheme in Algorithm 1 generates k samples and a linear combination of whose distributions approximates the posterior. In applications where it is desired to still generate one sample (rather than using a linear combination), we may consider a rejection sampling algorithm based on Theorem 3 to sample from $\tilde{P}_{X^n}(x|y = y^*)$. Let $\mathbf{T} = (T_1, \dots, T_m)^\top$ where $T_j = \sum_{i=1}^n \mathbb{1}_{X_i=u_j}$ for $j = 1, \dots, m$. Then $(g_1(\mathbf{T}/n), \dots, g_m(\mathbf{T}/n))^\top$ is the posterior probability vector associated with the plug-in posterior $\hat{P}_{X^n}(x|y = y^*)$ and $(D_{n,k}(g_1)(\mathbf{T}/n), \dots, D_{n,k}(g_m)(\mathbf{T}/n))^\top$ is the posterior probability vector associated with the debiased posterior $\tilde{P}_{X^n}(x|y = y^*)$. The rejection sampling is described in Algorithm 2.

Algorithm 2 Rejection Sampling for Debiased Posterior $\tilde{P}_{X^n}(x | y = y^*)$

Input: Plug-in posterior $\hat{P}_{X^n}(x | y = y^*)$, debiased posterior $\tilde{P}_{X^n}(x | y = y^*)$, large enough constant $M > 0$

Output: Sample from the debiased posterior $\tilde{P}_{X^n}(x | y = y^*)$

- 1: **repeat**
 - 2: Sample $x' \sim \hat{P}_{X^n}(x | y = y^*)$
 - 3: Sample $u \sim \text{Uniform}(0, M)$
 - 4: **until** $u < \frac{\tilde{P}_{X^n}(x' | y = y^*)}{\hat{P}_{X^n}(x' | y = y^*)}$
 - 5: **return** x'
-

In Algorithm 2,

$$M = \max_{x \in \mathcal{X}} \frac{\tilde{P}_{X^n}(x | y = y^*)}{\hat{P}_{X^n}(x | y = y^*)} = \max_j \left\{ \frac{D_{n,k}(g_j)(\mathbf{T}/n)}{g_j(\mathbf{T}/n)} \right\}$$

is the upper bound of the ratio of the debiased posterior to the plug-in posterior.

5 Experiments

In this section, we provide numerical experiments to illustrate the debiasing framework for posterior approximation under the binary prior case and the Gaussian mixture prior case.

Binary prior case. Suppose that $\mathcal{X} = \{0, 1\}$ and $X \sim \text{Bern}(q)$ for some unknown prior $q \in (0, 1)$. Let $\alpha = \alpha(y^*) := P_{Y|X}(y^*|1)/P_{Y|X}(y^*|0)$ be the likelihood ratio. Then the posterior distribution is give by $X|Y \sim \text{Bern}(\alpha q / (\alpha q + 1 - q))$. We estimate $g(q) := \alpha q / (\alpha q + 1 - q)$ based on the observations of X^n and y^* .

Proposition 1 provides a debiased approximation as a special case of Theorem 2 when $|\mathcal{X}| = 2$.

Proposition 1. Let $T = \sum_{i=1}^n X_i$. For $k = 1, 2, 3, 4$, we have

$$\mathbb{E}_{X^n} \{D_{n,k}g(T/n)\} - g(q) = \mathcal{O}(n^{-k}),$$

where $D_{n,k} = \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j B_n^j$ and $B_n(g)(x) = \sum_{k=0}^n g(\frac{k}{n}) \binom{n}{k} x^k (1-x)^{n-k}$ is the Bernstein polynomial approximation of g .

In the proof of Theorem 2, we notice that for any $k \in \mathbb{N}^+$, $\mathbb{E}_{X^n} \{D_{n,k}g(T/n)\} = C_{n,k}g(q)$, which allows Proposition 1 to be verified in closed form. To validate this result numerically, we consider two parameter settings: in the first experiment we set $q = 0.4$, $y^* = 2$, and $Y|X \sim \mathcal{N}(X, 1)$, while in the second we set $q = 3/11$, $y^* = 1$, and $Y|X \sim \mathcal{N}(X, 1/4)$.

For both settings, we examine the convergence rate of the debiased estimators $D_{n,k}g(T/n)$ for $k = 1, 2, 3, 4$. The results are shown in log-log plots in Figure 1, where the vertical axis represents the logarithm of the absolute error and the horizontal axis represents the logarithm of the sample size n . Reference lines with slopes corresponding to n^{-1} , n^{-2} , n^{-3} , and n^{-4} are included for comparison.

Gaussian mixture prior case. Suppose that $X \sim \frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{2}\mathcal{N}(1, 1)$ and $Y = X + \xi$ where $\xi \sim \mathcal{N}(0, 1/16)$. Additionally, let $y^* = 0.8$ and $A = \{x : x \geq 0.5\}$. In this case, we validate the theoretical convergence rate

$$|\mathbb{E}_{X^n} \{D_{n,k}f(\hat{\pi})(A)\} - f(\pi)(A)| = \mathcal{O}(n^{-k}).$$

Since $\mathbb{E}_{X^n} \{D_{n,k}f(\hat{\pi})(A)\}$ does not have a closed-form expression, we approximate it using Monte Carlo simulation. To ensure that the Monte Carlo error is negligible compared to the bias $\mathcal{O}(n^{-k})$, we select the number of Monte Carlo samples N such that $N \gg n^{2k-1}$. In practice, we run simulations for $k = 1$ and $k = 2$ and set $N = n^3$ for $k = 1$ and $N = n^4$ for $k = 2$.

The results are shown in Figure 2. The figure presents log-log plots where the vertical axis represents the logarithm of the absolute error or of the variance and the horizontal axis represents the logarithm of the sample size n . For both $k = 1$ and $k = 2$, the observed convergence rates align closely with the theoretical predictions.

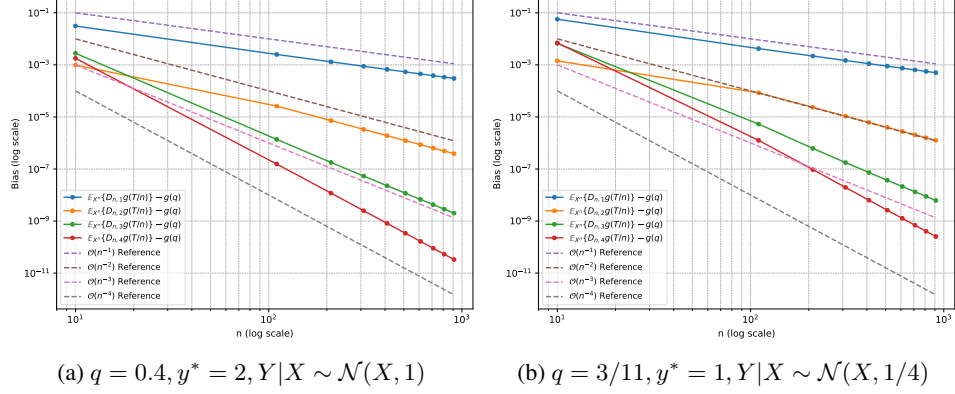


Figure 1: Convergence of plug-in and debiased estimators in the binary prior case. The plot compares the approximation error of $D_{n,k}g(T/n)$ ($k = 1, 2, 3, 4$) against n . Reference lines with slopes corresponding to n^{-1}, n^{-2}, n^{-3} , and n^{-4} are included to highlight the convergence rates.

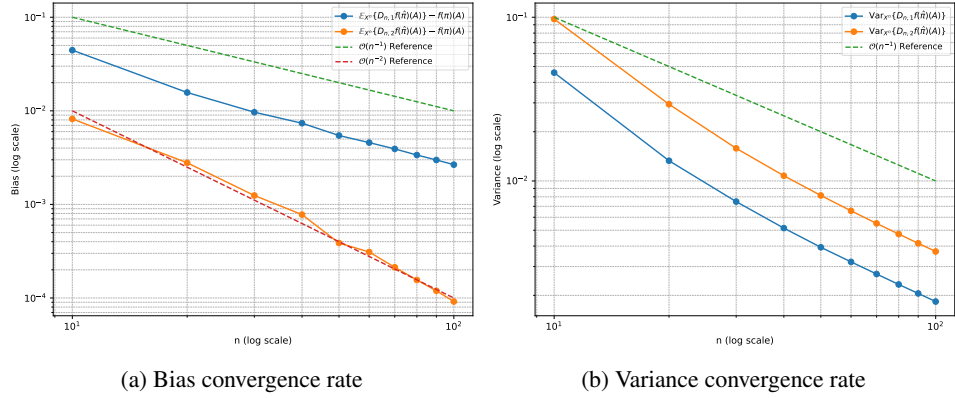


Figure 2: Convergence of debiased estimators in the Gaussian mixture prior case with $X \sim \frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{2}\mathcal{N}(1, 1)$, $Y = X + \xi$, $\xi \sim \mathcal{N}(0, 1/16)$, $y^* = 0.8$, and $A = \{x : x \geq 0.5\}$. (a) shows the bias decay of $D_{n,k}f(\hat{\pi})(A)$ for $k = 1, 2$, with reference lines of slopes corresponding to n^{-1} and n^{-2} included for comparison. (b) shows the corresponding variance decay, alongside a reference slope corresponding to n^{-1} .

6 Conclusion

We introduced a general framework for constructing a debiased posterior approximation through observed samples D and the known likelihood $P_{Y|X}$ when the prior distribution is unknown. Here, a naive strategy that directly plugs the empirical distribution into the Bayes formula or a generative model has a bias, because the likelihood is nonconstant, inducing a distribution shift, and the map from the prior to posterior is nonlinear. It can be shown that the plug-in approach generates $\hat{X}$ with bias $\|\mathbb{E}_D(P_{\hat{X}|Y=y^*,D}) - P_{X|Y=y^*}\|_{\text{TV}} = \mathcal{O}(n^{-1})$ and variance $\text{Var}_D(P_{\hat{X}|Y=y^*,D}) = \mathcal{O}(n^{-1})$. In contrast, our proposed debiasing framework achieves arbitrarily high-order bias rate of $\mathcal{O}(n^{-k})$ for any integer k , while maintaining the order of magnitude of the variance. Our framework is black-box in the sense that we only need to resample the training data and feed it into a given black-box conditional generative model. In particular, we provide a rigorous proof for the Bayes-optimal sampler f under the continuous prior setting and for a broad class of samplers f with a continuous $2k$ -th derivative, as specified in Assumption 2, under the discrete prior setting. We expect that the proposed debiasing framework could work for general f and will support future developments in bias-corrected posterior estimation and conditional sampling.

Acknowledgments

This research was supported in part by NSF Grant DMS-2515510.

References

- [1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations*, 2023.
- [2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [3] T. Tony Cai and Mark G. Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.
- [4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pages 5253–5270, 2023.
- [5] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.
- [6] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.
- [7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.
- [8] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915, 2022.
- [9] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.
- [10] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.
- [11] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.
- [12] Qiang Liu and Jason Lee. Black-box Importance Sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 952–961, 2017.
- [13] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.
- [14] Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International Conference on Learning Representations*, 2018.
- [15] Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control Functionals for Monte Carlo Integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3): 695–718, 05 2016.

- [16] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582, 2023.
- [17] Maurice Henry Quenouille. Notes on bias in estimation. *Biometrika*, 43(3-4):353–360, 1956.
- [18] Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont. Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 44565–44602, 2024.
- [19] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [20] Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2024.
- [21] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.
- [22] Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4958–4991, 2024.
- [23] Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, 17(4):402–586, 2020.
- [24] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.
- [25] TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.
- [26] Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Advances in Neural Information Processing Systems*, volume 36, pages 60599–60635, 2023.
- [27] Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortized inference. *Annual Review of Statistics and Its Application*, 12, 2024.
- [28] Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *International Conference on Machine Learning*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The model setting and the assumptions of our claim are clearly stated in abstract and introduction. Our main contributions are stated in introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We provide a black-box debiasing framework since we only provide the rigorous proof for the Bayes-optimal sampler under continuous prior setting and a broad class of samplers with a continuous $2k$ -th derivative under discrete prior setting. We expect the framework could work for general samplers f .

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All assumptions are clearly stated or referenced in the statement of our theorems and lemmas. The proofs appear in the main paper and the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We describe our experiments in the Experiment Section in detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We used the simulation data in the experiment.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the setting of our simulation in the Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiments do not include any error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Our experiment is a simple simulation for binary case.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our main contribution is constructing a debiased approximation of the posterior distribution which does not have immediate societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Proof of Theorem 1

Proof of Theorem 1. We begin by introducing notations that facilitates the analysis. Define

$$\mu = \int_{\mathcal{X}} \ell(x) \pi(dx), \quad \mu_A = \int_A \ell(x) \pi(dx).$$

Let $\hat{\pi}^{(0)} = \pi$, $\hat{\pi}^{(1)} = \hat{\pi}$ and set $(X_1^{(0)}, \dots, X_n^{(0)}) \equiv (X_1, \dots, X_n)$. For $j = 1, \dots, k$, we define $\hat{\pi}^{(j)}$ as the empirical measure of n i.i.d. samples $(X_1^{(j-1)}, \dots, X_n^{(j-1)})$ drawn from the measure $\hat{\pi}^{(j-1)}$. Furthermore, for each $j = 0, \dots, k$, define

$$e_n^{(j)} = n^{-1} \sum_{i=1}^n \left\{ \ell(X_i^{(j)}) - \mu \right\}, \quad \mu_A^{(j)} = n^{-1} \sum_{i=1}^n \ell(X_i^{(j)}) \delta_{X_i^{(j)}}(A).$$

Let

$$C_{n,k} = \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^j,$$

so that it suffices to show that

$$C_{n,k} f(\pi)(A) - f(\pi)(A) = \mathcal{O}_{L_1, L_2, k}(n^{-k}) \quad (9)$$

since $B_n D_{n,k} = C_{n,k}$.

The Radon-Nikodym derivative of $f(\pi)$ with respect to π is

$$\frac{df(\pi)}{d\pi}(x) = \frac{\ell(x)}{\int_{\mathcal{X}} \ell(x) \pi(dx)}.$$

For the empirical measure $\hat{\pi}$, the corresponding Radon-Nikodym derivative of $f(\hat{\pi})$ with respect to $\hat{\pi}$ takes the form

$$\begin{aligned} \frac{df(\hat{\pi})}{d\hat{\pi}}(x) &= \begin{cases} \frac{\ell(x)}{\int_{\mathcal{X}} \ell(x) \hat{\pi}(dx)}, & \text{if } x \in \{X_1^{(0)}, \dots, X_n^{(0)}\}, \\ 0, & \text{otherwise,} \end{cases} \\ &= \begin{cases} \frac{\ell(x)}{n^{-1} \sum_{i=1}^n \ell(X_i^{(0)})}, & \text{if } x \in \{X_1^{(0)}, \dots, X_n^{(0)}\}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently,

$$\begin{aligned} B_n f(\pi)(A) &= \mathbb{E} \{ f(\hat{\pi})(A) \} \\ &= \mathbb{E} \left\{ \int_A \frac{df(\hat{\pi})}{d\hat{\pi}}(x) \hat{\pi}(dx) \right\} \\ &= \mathbb{E} \left\{ \int_A \frac{\ell(x)}{n^{-1} \sum_{i=1}^n \ell(X_i^{(0)})} \hat{\pi}(dx) \right\} \\ &= \mathbb{E} \left\{ \frac{n^{-1} \sum_{i=1}^n \ell(X_i^{(0)}) \delta_{X_i^{(0)}}(A)}{n^{-1} \sum_{i=1}^n \ell(X_i^{(0)})} \right\}. \end{aligned}$$

Moreover, by the definition of B_n and iterated conditioning, we have $\mathbb{E} \{ f(\hat{\pi}^{(j)})(A) \} = \mathbb{E} [\mathbb{E} \{ f(\hat{\pi}^{(j)})(A) | \hat{\pi}^{(j-1)} \}] = \mathbb{E} \{ B_n f(\hat{\pi}^{(j-1)})(A) \} = \dots = \mathbb{E} \{ B_n^{j-1} f(\hat{\pi}^{(1)})(A) \} = B_n^j f(\pi)(A).$

By the same logic, for any $j = 2, \dots, k$, we have

$$\mathbb{E} \left\{ f(\hat{\pi}^{(j)})(A) \right\} = \mathbb{E} \left\{ \frac{n^{-1} \sum_{i=1}^n \ell(X_i^{(j-1)}) \delta_{X_i^{(j-1)}}(A)}{n^{-1} \sum_{i=1}^n \ell(X_i^{(j-1)})} \right\}.$$

Thus,

$$\begin{aligned} C_{n,k} f(\pi)(A) &= \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B_n^j f(\pi)(A) \\ &= \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left\{ f(\hat{\pi}^{(j)})(A) \right\} \\ &= \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left\{ \frac{n^{-1} \sum_{i=1}^n \ell(X_i^{(j-1)}) \delta_{X_i^{(j-1)}}(A)}{n^{-1} \sum_{i=1}^n \ell(X_i^{(j-1)})} \right\} \\ &= \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left(\frac{\mu_A^{(j-1)}}{e_n^{(j-1)} + \mu} \right). \end{aligned}$$

Then (9) holds if

$$\sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left(\frac{\mu_A^{(j-1)}}{e_n^{(j-1)} + \mu} \right) = \frac{\mu_A}{\mu} + \mathcal{O}_{L_1, L_2, k}(n^{-k}). \quad (10)$$

Now we show that (10) holds. By using the Taylor expansion of $1/(e_n^{(j-1)} + \mu)$, we have

$$\frac{1}{e_n^{(j-1)} + \mu} = \frac{1}{\mu} + \sum_{r=1}^{2k-1} \frac{(-1)^r}{\mu^{r+1}} (e_n^{(j-1)})^r + \frac{(e_n^{(j-1)})^{2k}}{\xi^{2k+1}},$$

where ξ lies between $e_n^{(j-1)} + \mu$ and μ .

Since $\min\{e_n^{(j-1)} + \mu, \mu\} \geq L_1$, we have $1/\xi^{2k+1} \leq L_1^{-2k-1}$. Additionally, $L_1 \leq l(X_i^{(j-1)}) \leq L_2$ and Hoeffding's inequality implies that

$$\mathbb{P}(|ne_n^{(j-1)}| > t) \leq 2 \exp \left\{ -\frac{2t^2}{n(L_2 - L_1)^2} \right\}$$

for all $t > 0$, which is equivalent to

$$\mathbb{P}(|e_n^{(j-1)}| > t) \leq 2 \exp \left\{ -\frac{2nt^2}{(L_2 - L_1)^2} \right\}.$$

Then

$$\begin{aligned} \mathbb{E} \left(|e_n^{(j-1)}|^{2k} \right) &= \int_0^\infty \mathbb{P} \left(|e_n^{(j-1)}|^{2k} > t \right) dt \\ &= \int_0^\infty \mathbb{P} \left(|e_n^{(j-1)}| > t^{1/2k} \right) dt \\ &\leq \int_0^\infty 2ku^{k-1} \exp \left\{ -\frac{2nu}{(L_2 - L_1)^2} \right\} du \\ &= 2kn^{-k} \int_0^\infty \exp \left\{ -\frac{2v}{(L_2 - L_1)^2} \right\} v^{k-1} dv \\ &= \mathcal{O}_{L_1, L_2, k}(n^{-k}). \end{aligned}$$

Therefore, we have

$$\mathbb{E} \left(\frac{\mu_A^{(j-1)}}{e_n^{(j-1)} + \mu} \right) = \frac{\mu_A}{\mu} + \mathbb{E} \left\{ \sum_{r=1}^{2k-1} \frac{(-1)^r}{\mu^{r+1}} \mu_A^{(j-1)} (e_n^{(j-1)})^r \right\} + \mathcal{O}_{L_1, L_2, k}(n^{-k}),$$

which implies that the L.H.S. of (10) can be written as

$$\frac{\mu_A}{\mu} + \sum_{r=1}^{2k-1} \frac{(-1)^r}{\mu^{r+1}} \left[\sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left\{ \mu_A^{(j-1)} (e_n^{(j-1)})^r \right\} \right] + \mathcal{O}_{L_1, L_2, k}(n^{-k}).$$

Thus to prove the bound on the bias, it remains to show that for any given k and $1 \leq r \leq 2k-1$,

$$B_{k,r} := \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \mathbb{E} \left\{ \mu_A^{(j-1)} (e_n^{(j-1)})^r \right\} = \mathcal{O}_{L_1, L_2, k}(n^{-k}).$$

Define a new operator $B : h \mapsto \mathbb{E}\{h(\hat{\pi})\}$ for any $h : \Delta_{\mathcal{X}} \rightarrow \mathbb{R}$ and let

$$h_s(\pi) = \left\{ \int_A \ell(x) \pi(dx) \right\} \left\{ \int \ell(x) \pi(dx) \right\}^s.$$

Since $B^j h_s(\pi) = \mathbb{E} \{ h_s(\hat{\pi}^{(j)}) \}$, we have

$$\begin{aligned} B_{k,r} &= \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} \sum_{s=0}^r \binom{r}{s} B^j h_s(\pi) (-1)^{(r-s)} \mu^{r-s} \\ &= \sum_{s=0}^r \binom{r}{s} (-1)^{r-s} \mu^{r-s} \sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B^j h_s(\pi). \end{aligned}$$

We claim that $B_{k,r} = \mathcal{O}_{L_1, L_2, k}(n^{-k})$ holds if for any $0 \leq s \leq r \leq 2k-1$,

$$(I - B)^k h_s(\pi) = \mathcal{O}_{L_1, L_2, s}(n^{-k}). \quad (11)$$

Indeed, $(I - B)^k h_s(\pi) = \mathcal{O}_{L_1, L_2, s}(n^{-k})$ is equivalent to $\sum_{j=1}^k \binom{k}{j} (-1)^{j-1} B^j h_s(\pi) = h_s(\pi) + \mathcal{O}_{L_1, L_2, s}(n^{-k})$. Therefore (11) implies

$$\begin{aligned} B_{k,r} &= \sum_{s=0}^r \binom{r}{s} (-1)^{r-s} \mu^{r-s} \{ h_s(\pi) + \mathcal{O}_{L_1, L_2, s}(n^{-k}) \} \\ &= \sum_{s=0}^r \left\{ \binom{r}{s} (-1)^{r-s} \mu_A \mu^r + \mathcal{O}_{L_1, L_2, s}(n^{-k}) \right\} \\ &= \mathcal{O}_{L_1, L_2, k}(n^{-k}). \end{aligned}$$

Now, to prove the bound on the bias we only need to show that (11) holds. For any $k \in \mathbb{N}$ and $s \in \mathbb{N}^+$, let

$$\mathfrak{J}_s := \left\{ (\mathbf{a}, \mathbf{s}, v) : \mathbf{a} = (a_1, a_2, \dots), \mathbf{s} = (s_1, s_2, \dots), a_i, s_i \in \mathbb{N}^+, v \in \mathbb{N}, a_1 > a_2 > \dots \geq 1, \sum_i a_i s_i + v = s \right\}$$

and

$$\mathcal{A}_s^k := \left\{ \sum_{(\mathbf{a}, \mathbf{s}, v) \in \mathfrak{J}_s} \alpha_{\mathbf{a}, \mathbf{s}, v} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} : |\alpha_{\mathbf{a}, \mathbf{s}, v}| \leq C_k(s) n^{-k} \right\},$$

where $C_0(s), C_1(s), \dots$ are constants from Lemma 2. Since $h_s(\pi) \in \mathcal{A}_{s+1}^0$, Lemma 2 implies that $(I - B)^k h_s(\pi) \in \mathcal{A}_{s+1}^k$. Therefore, $(I - B)^k h_s(\pi) = \mathcal{O}_{L_1, L_2, s}(n^{-k})$, finishing the proof for the bias bound.

Finally, to prove the bound on the variance, consider the function $F(x, y) = x/y$. By construction,

$$f(\hat{\pi}^{(j)})(A) = F(\mu_A^{(j-1)}, e_n^{(j-1)} + \mu).$$

Applying the Taylor expansion of $F(x, y)$ yields

$$f(\hat{\pi}^{(j)})(A) = f(\pi)(A) + \frac{1}{\mu_A}(\mu_A^{(j-1)} - \mu_A) - \frac{\mu_A}{\mu^2}e_n^{(j-1)} - \frac{1}{\xi_y^2}(\mu_A^{(j-1)} - \mu_A)e_n^{(j-1)} + \frac{\xi_x}{\xi_y^3}(e_n^{(j-1)})^2,$$

for some ξ_x lying between μ_A and $\mu_A^{(j-1)}$, and ξ_y lying between μ and $e_n^{(j-1)} + \mu$. Since $L_1 \leq \mu_A, \mu_A^{(j-1)}, \mu, e_n^{(j-1)} + \mu \leq L_2$ implies that $|1/\xi_y^2|$ and $|\xi_x/\xi_y^3|$ are bounded by some constant depending on L_1 and L_2 .

Moreover, since

$$\begin{aligned} \left| \mathbb{E} \left\{ (\mu_A^{(j-1)} - \mu_A) e_n^{(j-1)} \right\} \right| &= \left| \text{Cov} \left(\mu_A^{(j-1)}, e_n^{(j-1)} + \mu \right) \right| \\ &\leq \left\{ \text{Var} \left(\mu_A^{(j-1)} \right) \right\}^{1/2} \left\{ \text{Var} \left(e_n^{(j-1)} + \mu \right) \right\}^{1/2} \\ &= \left[\text{Var} \left\{ n^{-1} \sum_{i=1}^n \ell(X_i^{(j)}) \delta_{X_i^{(j-1)}}(A) \right\} \right]^{1/2} \left[\text{Var} \left\{ n^{-1} \sum_{i=1}^n \ell(X_i^{(j-1)}) \right\} \right]^{1/2} \\ &= \left[\frac{1}{n} \text{Var} \left\{ \ell(X_i^{(j)}) \delta_{X_i^{(j-1)}}(A) \right\} \right]^{1/2} \left[\frac{1}{n} \text{Var} \left\{ \ell(X_i^{(j-1)}) \right\} \right]^{1/2} \\ &= \mathcal{O}_{L_1, L_2}(n^{-1}), \end{aligned}$$

and

$$\mathbb{E} \left(e_n^{(j-1)} \right)^2 = \frac{1}{n} \text{Var} \left\{ \ell(X_i^{(j-1)}) \right\} = \mathcal{O}_{L_1, L_2}(n^{-1}).$$

Combining these bounds with the Taylor expansion, we conclude that for any $j \geq 1$,

$$B_n^j f(\pi)(A) = \mathbb{E} \left\{ f(\hat{\pi}^{(j)})(A) \right\} = f(\pi)(A) + \mathcal{O}_{L_1, L_2}(n^{-1}).$$

By the same logic, we also have $B_n \{f(\pi)(A)\}^2 = \{f(\pi)(A)\}^2 + \mathcal{O}_{L_1, L_2}(n^{-1})$.

Therefore,

$$\begin{aligned} D_{n,k} f(\pi)(A) &= \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j B_n^j f(\pi)(A) \\ &= \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j \{f(\pi)(A) + \mathcal{O}_{L_1, L_2}(n^{-1})\} \\ &= f(\pi)(A) + \mathcal{O}_{L_1, L_2, k}(n^{-1}), \end{aligned}$$

and

$$\begin{aligned} \text{Var}_{X^n} \{D_{n,k} f(\hat{\pi})(A)\} &= \mathbb{E} \left[\{D_{n,k} f(\hat{\pi})(A)\}^2 \right] - [\mathbb{E} \{D_{n,k} f(\hat{\pi})(A)\}]^2 \\ &= B_n \{D_{n,k} f(\pi)(A)\}^2 - \{f(\pi)(A) + \mathcal{O}_{L_1, L_2, k}(n^{-1})\}^2 \\ &= B_n \{f(\pi)(A) + \mathcal{O}_{L_1, L_2, k}(n^{-1})\}^2 - \{f(\pi)(A) + \mathcal{O}_{L_1, L_2, k}(n^{-1})\}^2 \\ &= B_n \{f(\pi)(A)\}^2 + \mathcal{O}_{L_1, L_2, k}(n^{-1}) - \{f(\pi)(A)\}^2 \\ &= \mathcal{O}_{L_1, L_2, k}(n^{-1}). \end{aligned}$$

□

Lemma 2. *There exist constants $C_0(s), C_1(s), C_2(s), \dots$, such that the following holds.*

For any $k \in \mathbb{N}$ and $s, n \in \mathbb{N}^+$, let

$$\mathfrak{J}_s := \left\{ (\mathbf{a}, \mathbf{s}, v) : \mathbf{a} = (a_1, a_2, \dots), \mathbf{s} = (s_1, s_2, \dots), a_i, s_i \in \mathbb{N}^+, v \in \mathbb{N}, a_1 > a_2 > \dots \geq 1, \sum_i a_i s_i + v = s \right\}$$

and

$$\mathcal{A}_s^k := \left\{ \sum_{(\mathbf{a}, \mathbf{s}, v) \in \mathfrak{J}_s} \alpha_{\mathbf{a}, \mathbf{s}, v} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} : |\alpha_{\mathbf{a}, \mathbf{s}, v}| \leq C_k(s) n^{-k} \right\}.$$

If $h(\pi) \in \mathcal{A}_s^0$, then for any $k \in \mathbb{N}$, we have

$$(I - B)^k h(\pi) \in \mathcal{A}_s^k, \quad (12)$$

where B is an operator defined as $Bh(\pi) = \mathbb{E}\{h(\hat{\pi})\}$ where $\hat{\pi}$ is the empirical distribution of $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \pi$.

Proof of Lemma 2. We begin by proving that $(I - B)h(\pi) \in \mathcal{A}_s^1$. Since $h(\pi) \in \mathcal{A}_s^0$, let

$$h(\pi) = \sum_{(\mathbf{a}, \mathbf{s}, v) \in \mathfrak{J}_s} \alpha_{\mathbf{a}, \mathbf{s}, v} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i}.$$

Note that $|\mathfrak{J}_s|$ does not depend on n and $|\alpha_{\mathbf{a}, \mathbf{s}, v}| \leq C_0(s)$. It suffices to verify that each individual term in the sum satisfies

$$(I - B) \left[\left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \right] \in \mathcal{A}_s^1.$$

Without loss of generality, let $\mathbf{a} = (a_1, \dots, a_p)$ and $\mathbf{s} = (s_1, \dots, s_p)$, $s' = \sum_{i=1}^p s_i$. Then we have $\sum_i a_i s_i + v = s$ and

$$\begin{aligned} & B \left[\left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \right] \\ &= \mathbb{E} \left[\left\{ \int_A \ell^v(x) \hat{\pi}(dx) \right\} \prod_{i=1}^p \left\{ \int \ell^{a_i}(x) \hat{\pi}(dx) \right\}^{s_i} \right] \\ &= \frac{1}{n^{s'+1}} \mathbb{E} \left[\left\{ \sum_{j=1}^n \ell^v(X_j) \delta_{X_j}(A) \right\} \prod_{i=1}^p \left\{ \sum_{j=1}^n \ell^{a_i}(X_j) \right\}^{s_i} \right]. \end{aligned}$$

For the term $\prod_{i=1}^p \left\{ \sum_{j=1}^n \ell^{a_i}(X_j) \right\}^{s_i}$, let $m_j^{(i)}$ denote the times X_j appears with powers a_i , then we have $\sum_{j=1}^n m_j^{(i)} = s_i$ for $1 \leq i \leq p$. Define

$$\mathcal{I}_s = \left\{ \mathbf{m} = (m_j^{(i)})_{j \in [n], i \in [p]} : \sum_{j=1}^n m_j^{(i)} = s_i \text{ for all } i \in [p] \right\}.$$

Therefore,

$$\left\{ \sum_{j=1}^n \ell^v(X_j) \delta_{X_j}(A) \right\} \prod_{i=1}^p \left\{ \sum_{j=1}^n \ell^{a_i}(X_j) \right\}^{s_i} = \left\{ \sum_{j=1}^n \ell^v(X_j) \delta_{X_j}(A) \right\} \sum_{\mathbf{m} \in \mathcal{I}_s} c_{\mathbf{s}, \mathbf{m}} \prod_{j=1}^n \ell^{\sum_{i=1}^p a_i m_j^{(i)}}(X_j), \quad (13)$$

where

$$c_{\mathbf{s}, \mathbf{m}} = \prod_{i=1}^p \frac{s_i!}{\prod_{j=1}^n m_j^{(i)}!}.$$

Note that $c_{\mathbf{s}, \mathbf{m}}$ does not depend on n . Now we expand R.H.S. of (13) based on the number of distinct variables X_j appear, i.e., $\sum_{j=1}^n \mathbb{1}_{\sum_{i=1}^p a_i m_j^{(i)} > 0}$, which is equal to $\sum_{j=1}^n \mathbb{1}_{\sum_{i=1}^p m_j^{(i)} > 0}$. Define

$$\mathcal{J}_{\mathbf{m}} = \left\{ j \in [n] : \sum_{i=1}^p m_j^{(i)} > 0 \right\},$$

then we have $1 \leq |\mathcal{J}_{\mathbf{m}}| \leq s'$.

Hence,

$$\begin{aligned} & \mathbb{E} \left[\left\{ \sum_{j=1}^n \ell^v(X_j) \delta_{X_j}(A) \right\} \prod_{i=1}^p \left\{ \sum_{j=1}^n \ell^{a_i}(X_j) \right\}^{s_i} \right] \\ &= \mathbb{E} \left[\left\{ \sum_{j=1}^n \ell^v(X_j) \delta_{X_j}(A) \right\} \sum_{m=1}^{s'} \sum_{\substack{\mathbf{m} \in \mathcal{I}_{\mathbf{s}} \\ |\mathcal{J}_{\mathbf{m}}|=m}} c_{\mathbf{s}, \mathbf{m}} \prod_{j=1}^n \ell^{\sum_{i=1}^p a_i m_j^{(i)}}(X_j) \right] \\ &= \mathbb{E} \left\{ \sum_{m=1}^{s'} \sum_{\substack{\mathbf{m} \in \mathcal{I}_{\mathbf{s}} \\ |\mathcal{J}_{\mathbf{m}}|=m}} c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A) \prod_{j=1}^n \ell^{\sum_{i=1}^p a_i m_j^{(i)}}(X_j) \right\} \\ &= n(n-1) \cdots (n-s') c_{\mathbf{s}, \mathbf{m}^*} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p [\mathbb{E} \{\ell^{a_i}(X)\}]^{s_i} \\ &\quad + \mathbb{E} \left\{ \sum_{\substack{\mathbf{m} \in \mathcal{I}_{\mathbf{s}} \\ |\mathcal{J}_{\mathbf{m}}|=s'}} c_{\mathbf{s}, \mathbf{m}} \sum_{t \in \mathcal{J}_{\mathbf{m}}} \ell^v(X_t) \delta_{X_t}(A) \prod_{j=1}^n \ell^{\sum_{i=1}^p a_i m_j^{(i)}}(X_j) \right\} \\ &\quad + \mathbb{E} \left\{ \sum_{m=1}^{s'-1} \sum_{\substack{\mathbf{m} \in \mathcal{I}_{\mathbf{s}} \\ |\mathcal{J}_{\mathbf{m}}|=m}} c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A) \prod_{j=1}^n \ell^{\sum_{i=1}^p a_i m_j^{(i)}}(X_j) \right\}, \end{aligned} \quad (14)$$

where $c_{\mathbf{s}, \mathbf{m}^*} = \prod_{i=1}^p s_i!$.

The three terms in (14) are interpreted as follows: we can expand $\{\sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A)\} \prod_{i=1}^p \left\{ \sum_{j=1}^n \ell^{a_i}(X_j) \right\}^{s_i}$ as the sum of many product terms of the form $\ell^v(X_t) \prod_{i=1}^p \prod_{l=1}^{s_i} \ell^{a_i}(X_{j_{i,l}})$. The first term in (14) corresponds to the partial sum of terms in which all of $X_t, (X_{j_{i,l}})_{i,l}$ are distinct. The second term in (14) corresponds to the partial sum of terms in which X_t is identical to one of $(X_{j_{i,l}})_{i,l}$ while the latter are distinct. The third term corresponds to the partial sum of terms in which at least two of $(X_{j_{i,l}})_{i,l}$ are identical. The last two term in (14) are at least $\mathcal{O}(n^{-1})$ factor smaller than the first (due to fewer terms involved in the sum because of the constraint of having identical terms), while the first term will cancel with $I \cdot h(\pi)$ when applying $I - B$ to h .

Let $\mathcal{P}(b_1, \dots, b_m)$ denote the set of all distinct permutations of the vector consisting of m non-zero elements $b_1, \dots, b_m$ and $n - m$ zeros. Note that even the values of b_i may be the same, we still treat the b_i s are distinguishable. Then since 0s are identical, we have $|\mathcal{P}(b_1, \dots, b_m)| = n(n-1) \cdots (n-m+1) = \mathcal{O}(n^m)$. Additionally, for any $\mathbf{a}$ and $\mathbf{m}$, we define

$$\Psi(\mathbf{a}, \mathbf{m}) = \left(\sum_{i=1}^p a_i m_1^{(i)}, \dots, \sum_{i=1}^p a_i m_n^{(i)} \right).$$

Now we can write (14) as

$$n(n-1) \cdots (n-s') c_{\mathbf{s}, \mathbf{m}^*} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p [\mathbb{E} \{\ell^{a_i}(X)\}]^{s_i}$$

$$\begin{aligned}
& + \sum_{b_k: \sum_{k=1}^{s'} b_k = s-v} \sum_{\mathbf{m}: \Psi(\mathbf{a}, \mathbf{m}) \in \mathcal{P}(b_1, \dots, b_{s'})} c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^m \left[\prod_{i \neq t} \mathbb{E}\{\ell^{b_i}(X)\} \right] \int_A \ell^{b_t+v}(x) \pi(dx) \\
& + \sum_{m=1}^{s'-1} \sum_{b_k: \sum_{k=1}^m b_k = s-v} \sum_{\mathbf{m}: \Psi(\mathbf{a}, \mathbf{m}) \in \mathcal{P}(b_1, \dots, b_m)} c_{\mathbf{s}, \mathbf{m}} \mathbb{E} \left\{ \sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A) \prod_{i=1}^m \ell^{b_i}(X_i) \right\} \\
& = n(n-1) \cdots (n-s') c_{\mathbf{s}, \mathbf{m}^*} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p [\mathbb{E}\{\ell^{a_i}(X)\}]^{s_i} \\
& + \sum_{b_k: \sum_{k=1}^{s'} b_k = s-v} \mathcal{O}(n^{s'}) c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^m \left[\prod_{i \neq t} \mathbb{E}\{\ell^{b_i}(X)\} \right] \int_A \ell^{b_t+v}(x) \pi(dx) \\
& + \sum_{m=1}^{s'-1} \sum_{b_k: \sum_{k=1}^m b_k = s-v} \mathcal{O}(n^m) c_{\mathbf{s}, \mathbf{m}} \mathbb{E} \left\{ \sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A) \prod_{i=1}^m \ell^{b_i}(X_i) \right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& (I - B) \left[\left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \right] \\
& = \frac{n^{s'+1} - n(n-1) \cdots (n-s')}{n^{s'+1}} c_{\mathbf{s}, \mathbf{m}^*} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p [\mathbb{E}\{\ell^{a_i}(X)\}]^{s_i} \\
& - \sum_{b_k: \sum_{k=1}^{s'} b_k = s-v} \mathcal{O}(n^{-1}) c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^m \left[\prod_{i \neq t} \mathbb{E}\{\ell^{b_i}(X)\} \right] \int_A \ell^{b_t+v}(x) \pi(dx) \\
& - \sum_{m=1}^{s'-1} \sum_{b_k: \sum_{k=1}^m b_k = s-v} \mathcal{O}(n^{m-s'-1}) c_{\mathbf{s}, \mathbf{m}} \mathbb{E} \left\{ \sum_{t=1}^n \ell^v(X_t) \delta_{X_t}(A) \prod_{i=1}^m \ell^{b_i}(X_i) \right\} \\
& = \frac{n^{s'+1} - n(n-1) \cdots (n-s')}{n^{s'+1}} c_{\mathbf{s}, \mathbf{m}^*} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^p [\mathbb{E}\{\ell^{a_i}(X)\}]^{s_i} \\
& - \sum_{b_k: \sum_{k=1}^{s'} b_k = s-v} \mathcal{O}(n^{-1}) c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^m \left[\prod_{i \neq t} \mathbb{E}\{\ell^{b_i}(X)\} \right] \int_A \ell^{b_t+v}(x) \pi(dx) \\
& - \sum_{m=1}^{s'-1} \sum_{b_k: \sum_{k=1}^m b_k = s-v} \mathcal{O}(n^{m-s'-1}) c_{\mathbf{s}, \mathbf{m}} \sum_{t=1}^m \left[\prod_{i \neq t} \mathbb{E}\{\ell^{b_i}(X)\} \right] \int_A \ell^{b_t+v}(x) \pi(dx) \\
& - \sum_{m=1}^{s'-1} \sum_{b_k: \sum_{k=1}^m b_k = s-v} \mathcal{O}(n^{m-s'}) c_{\mathbf{s}, \mathbf{m}} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_{i=1}^m \mathbb{E}\{\ell^{b_i}(X_i)\} \\
& \in \mathcal{A}_s^1.
\end{aligned}$$

The last inclusion follows from that fact that $\{n^{s'+1} - n(n-1) \cdots (n-s')\}/n^{s'+1} = \mathcal{O}(n^{-1})$ and the number of solutions to $\sum_{k=1}^m b_k = s-v$ does not depend on n but depends on s .

Now we suppose (12) holds for k . Then we can set

$$(I - B)^k h(\pi) = \sum_{(\mathbf{a}, \mathbf{s}, v) \in \mathfrak{J}_s} \alpha'_{\mathbf{a}, \mathbf{s}, v} \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i},$$

where $|\alpha'_{\mathbf{a},\mathbf{s},v}| \leq C_k(s)n^{-k}$. Then for $k+1$, we have

$$(I-B)^{k+1}h(\pi) = \sum_{(\mathbf{a},\mathbf{s},v) \in \mathfrak{J}_s} \alpha'_{\mathbf{a},\mathbf{s},v} (I-B) \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i}.$$

Since for all $\mathbf{a}, \mathbf{s}, v$ such that $(\mathbf{a}, \mathbf{s}, v) \in \mathfrak{J}_s$, $\left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \in \mathcal{A}_s^0$, we have $(I-B) \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \in \mathcal{A}_s^1$, namely,

$$\begin{aligned} & (I-B) \left\{ \int_A \ell^v(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{a_i}(x) \pi(dx) \right\}^{s_i} \\ &= \sum_{(\mathbf{b},\mathbf{t},u) \in \mathfrak{J}_s} \alpha_{\mathbf{b},\mathbf{t},u}(\mathbf{a}, \mathbf{s}, v) \left\{ \int_A \ell^u(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{b_i}(x) \pi(dx) \right\}^{t_i}, \end{aligned}$$

where $|\alpha_{\mathbf{b},\mathbf{t},u}(\mathbf{a}, \mathbf{s}, v)| \leq C_0(s)n^{-1}$. Therefore,

$$\begin{aligned} & (I-B)^{k+1}h(\pi) \\ &= \sum_{(\mathbf{a},\mathbf{s},v) \in \mathfrak{J}_s} \alpha'_{\mathbf{a},\mathbf{s},v} \sum_{(\mathbf{b},\mathbf{t},u) \in \mathfrak{J}_s} \alpha_{\mathbf{b},\mathbf{t},u}(\mathbf{a}, \mathbf{s}, v) \left\{ \int_A \ell^u(x) \pi(dx) \right\} \prod_i \left\{ \int \ell^{b_i}(x) \pi(dx) \right\}^{t_i} \\ &\in \mathcal{A}_s^{k+1}. \end{aligned}$$

□

B Proof of Theorem 2

In order to prove Theorem 2, we first make some preliminary observations.

Let function f defined on the simplex $\Delta_m = \{\mathbf{q} \in \mathbb{R}^m : q_j \geq 0, \sum_{j=1}^m q_j = 1\}$. Define the generalized Bernstein basis polynomials of degree n as

$$b_{n,\nu}(\mathbf{q}) = \binom{n}{\nu} \mathbf{q}^\nu.$$

Lemma 3. $|\sum_{\nu \in \bar{\Delta}_m} (\nu/n - \mathbf{q})^\alpha b_{n,\nu}(\mathbf{q})| \lesssim n^{-\|\alpha\|_1/2}$.

Proof of Lemma 3. It suffices to show that $|\sum_{\nu \in \bar{\Delta}_m} (\nu - n\mathbf{q})^\alpha b_{n,\nu}(\mathbf{q})| \lesssim n^{\|\alpha\|_1/2}$. Since $\mathbf{q} \in \Delta_m$, we treat $T_{n,\alpha} \equiv \sum_{\nu \in \bar{\Delta}_m} (\nu - n\mathbf{q})^\alpha b_{n,\nu}(\mathbf{q})$ as a function of the variables $q_1, \dots, q_{m-1}$. For any $\beta \in \mathbb{N}^{m-1}$ such that $\|\beta\|_1 = 1$, we let $\gamma = \gamma(\beta) \equiv (\beta^\top, 0)^\top$. Additionally, let $\theta = (0, \dots, 0, 1)^\top \in \mathbb{N}^m$. Since

$$\partial^\beta (\nu - n\mathbf{q})^\alpha = -n\alpha^\gamma (\nu - n\mathbf{q})^{\alpha-\gamma} + n\alpha^\theta (\nu - n\mathbf{q})^{\alpha-\theta},$$

and

$$\begin{aligned} \partial^\beta b_{n,\nu}(\mathbf{q}) &= \binom{n}{\nu} (\nu^\gamma \mathbf{q}^{\nu-\gamma} - \nu^\theta \mathbf{q}^{\nu-\theta}) \\ &= b_{n,\nu}(\mathbf{q}) \left\{ \frac{1}{\mathbf{q}^\gamma} (\nu - n\mathbf{q})^\gamma - \frac{1}{\mathbf{q}^\theta} (\nu - n\mathbf{q})^\theta \right\}, \end{aligned}$$

we have

$$\begin{aligned} \partial^\beta T_{n,\alpha} &= \sum_{\nu \in \bar{\Delta}_m} \partial^\beta (\nu - n\mathbf{q})^\alpha b_{n,\nu}(\mathbf{q}) + \sum_{\nu \in \bar{\Delta}_m} (\nu - n\mathbf{q})^\alpha \partial^\beta b_{n,\nu}(\mathbf{q}) \\ &= -n\alpha^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta T_{n,\alpha-\theta} + \frac{1}{\mathbf{q}^\gamma} T_{n,\alpha+\gamma} - \frac{1}{\mathbf{q}^\theta} T_{n,\alpha+\theta}, \end{aligned}$$

i.e.,

$$\mathbf{q}^\gamma \partial^\beta T_{n,\alpha} = -n\alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta \mathbf{q}^\gamma T_{n,\alpha-\theta} + T_{n,\alpha+\gamma} - \frac{\mathbf{q}^\gamma}{\mathbf{q}^\theta} T_{n,\alpha+\theta}.$$

By summing the above equation over $\beta \in \mathbb{N}^{m-1}$ such that $\|\beta\|_1 = 1$, we have

$$\begin{aligned}
& \sum_{\|\beta\|_1=1} \mathbf{q}^\gamma \partial^\beta T_{n,\alpha} \\
&= -n \sum_{\|\beta\|_1=1} \alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta \sum_{\|\beta\|_1=1} \mathbf{q}^\gamma T_{n,\alpha-\theta} + \sum_{\|\beta\|_1=1} T_{n,\alpha+\gamma} - \frac{1-\mathbf{q}^\theta}{\mathbf{q}^\theta} T_{n,\alpha+\theta} \\
&= -n \sum_{\|\beta\|_1=1} \alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta (1-\mathbf{q}^\theta) T_{n,\alpha-\theta} + \sum_{\|\beta\|_1=1} T_{n,\alpha+\gamma} - \frac{1-\mathbf{q}^\theta}{\mathbf{q}^\theta} T_{n,\alpha+\theta} \\
&= -n \sum_{\|\beta\|_1=1} \alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta (1-\mathbf{q}^\theta) T_{n,\alpha-\theta} - \frac{1}{\mathbf{q}^\theta} T_{n,\alpha+\theta},
\end{aligned}$$

where the last equality follows from the fact that $\sum_{\|\beta\|_1=1} T_{n,\alpha+\gamma} + T_{n,\alpha+\theta} = 0$.

Therefore, we have the following recurrence formula:

$$T_{n,\alpha+\theta} = -n\mathbf{q}^\theta \sum_{\|\beta\|_1=1} \alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} + n\alpha^\theta \mathbf{q}^\theta (1-\mathbf{q}^\theta) T_{n,\alpha-\theta} - \mathbf{q}^\theta \sum_{\|\beta\|_1=1} \mathbf{q}^\gamma \partial^\beta T_{n,\alpha}, \quad (15)$$

$$T_{n,\alpha+\gamma} = \frac{\mathbf{q}^\gamma}{\mathbf{q}^\theta} T_{n,\alpha+\theta} + n\alpha^\gamma \mathbf{q}^\gamma T_{n,\alpha-\gamma} - n\alpha^\theta \mathbf{q}^\gamma T_{n,\alpha-\theta} + \mathbf{q}^\gamma \partial^\beta T_{n,\alpha}. \quad (16)$$

Using the recurrence fomrula (15), (16) and the fact that $T_{n,(1,0,\dots,0)^\top} = 0$, $T_{n,(2,0,\dots,0)^\top} = nq_1(1-q_1)$, $T_{n,(1,1,0,\dots,0)^\top} = -nq_1q_2$, $T_{n,\alpha}$ has the following form by using induction:

$$T_{n,\alpha} = \sum_{j=1}^{\lfloor \|\alpha\|_1/2 \rfloor} n^j \left(\sum_{\eta \leq \alpha} c_{j,\eta} \mathbf{q}^\eta \right), \quad (17)$$

where $c_{j,\eta}$ is independent of n . Then we can conclude that $|T_{n,\alpha}| \lesssim n^{\lfloor \|\alpha\|_1/2 \rfloor} \lesssim n^{\|\alpha\|_1/2}$.

□

Proof of Lemma 1. We prove the theorem by induction on k .

For $k = 1$, by Taylor's expansion, there exists $\xi \in \Delta_m$ such that

$$f\left(\frac{\nu}{n}\right) = f(\mathbf{q}) + \sum_{\|\alpha\|_1=1} \partial^\alpha f(\xi) \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha.$$

Then we have

$$\begin{aligned}
|B_n(f)(\mathbf{q}) - f(\mathbf{q})| &= \left| \sum_{\nu \in \bar{\Delta}_m} \left\{ f\left(\frac{\nu}{n}\right) - f(\mathbf{q}) \right\} b_{n,\nu}(\mathbf{q}) \right| \\
&= \left| \sum_{\nu \in \bar{\Delta}_m} \left\{ \sum_{\|\alpha\|_1=1} \partial^\alpha f(\xi) \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha \right\} b_{n,\nu}(\mathbf{q}) \right| \\
&\leq \|f\|_{C^1(\Delta_m)} \sum_{\|\alpha\|_1=1} \sum_{\nu \in \bar{\Delta}_m} \left| \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha \right| b_{n,\nu}(\mathbf{q}) \\
&\leq \|f\|_{C^1(\Delta_m)} \sum_{\|\alpha\|_1=1} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^{2\alpha} b_{n,\nu}(\mathbf{q}) \right\}^{1/2} \\
&\lesssim \|f\|_{C^1(\Delta_m)} \sum_{\|\alpha\|_1=1} n^{-1/2} \\
&\lesssim_m \|f\|_{C^1(\Delta_m)} n^{-1/2},
\end{aligned}$$

where the second inequality follows from Cauchy–Schwarz inequality, and the third inequality follows from Lemma 3.

Suppose the theorem holds up through k . Now we prove the theorem for $k+1$. For $k+1$, by Taylor's expansion, there exists $\xi \in \Delta_m$ such that

$$f\left(\frac{\nu}{n}\right) = f(\mathbf{q}) + \sum_{\|\alpha\|_1=1}^k \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha + \sum_{\|\alpha\|_1=k+1} \frac{\partial^\alpha f(\xi)}{\alpha!} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha.$$

Then we have

$$\begin{aligned} B_n(f)(\mathbf{q}) - f(\mathbf{q}) &= \sum_{\nu \in \bar{\Delta}_m} \left(f\left(\frac{\nu}{n}\right) - f(\mathbf{q}) \right) b_{n,\nu}(\mathbf{q}) \\ &= \sum_{\nu \in \bar{\Delta}_m} \left\{ \sum_{\|\alpha\|_1=1}^k \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha + \sum_{\|\alpha\|_1=k+1} \frac{\partial^\alpha f(\xi)}{\alpha!} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha \right\} b_{n,\nu}(\mathbf{q}) \\ &= \sum_{\|\alpha\|_1=1}^k \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \\ &\quad + \sum_{\|\alpha\|_1=k+1} \frac{\partial^\alpha f(\xi)}{\alpha!} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} (B_n - I)^{\lceil (k+1)/2 \rceil} (f)(\mathbf{q}) &= \sum_{\|\alpha\|_1=1}^k (B_n - I)^{\lceil (k+1)/2 \rceil - 1} \left\{ \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \\ &\quad + \sum_{\|\alpha\|_1=k+1} (B_n - I)^{\lceil (k+1)/2 \rceil - 1} \left\{ \frac{\partial^\alpha f(\xi)}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\}. \end{aligned} \tag{18}$$

First, we consider the first term of the right-hand side of (18). We know that $(\alpha!)^{-1} \partial^\alpha f(\mathbf{q}) \sum_{\nu \in \bar{\Delta}_m} (\nu/n - \mathbf{q})^\alpha b_{n,\nu}(\mathbf{q}) \in C^{k+1-\|\alpha\|_1}(\Delta_m)$ since $f \in C^{k+1}(\Delta_m)$. By the induction hypothesis, we have

$$\begin{aligned} &\left\| (B_n - I)^{\lceil (k+1-\|\alpha\|_1)/2 \rceil} \left\{ \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ &\lesssim_{k+1-\|\alpha\|_1, m} \left\| \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\|_{C^{k+1-\|\alpha\|_1}(\Delta_m)} n^{-(k+1-\|\alpha\|_1)/2}. \end{aligned}$$

Let

$$g_\alpha(\mathbf{q}) = \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}),$$

For any $|\beta| \leq k+1 - \|\alpha\|_1$, we have

$$\begin{aligned} \|\partial^\beta g_\alpha(\mathbf{q})\|_\infty &= \left\| \frac{1}{\alpha!} \sum_{0 \leq \gamma \leq \beta} \binom{\beta}{\gamma} \partial^{\alpha+\gamma} f(\mathbf{q}) \partial^{\beta-\gamma} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ &\lesssim_{k+1} \|f\|_{C^{k+1}(\Delta_m)} \sum_{0 \leq \gamma \leq \beta} \binom{\beta}{\gamma} \left\| \partial^{\beta-\gamma} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \end{aligned}$$

$$\lesssim_{k+1} \|f\|_{C^{k+1}(\Delta_m)} n^{-\|\alpha\|_1/2},$$

where the last inequality follows from the fact that $\|\partial^{\beta-\gamma} \{\sum_{\nu \in \bar{\Delta}_m} (\nu/n - \mathbf{q})^\alpha b_{n,\nu}(\mathbf{q})\}\|_\infty \lesssim n^{-\|\alpha\|_1/2}$ which can be derived by using the form of $T_{n,\alpha}$ in (17).

Therefore, we have

$$\begin{aligned} & \left\| (B_n - I)^{\lceil (k+1-\|\alpha\|_1)/2 \rceil} \left\{ \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ & \lesssim_{k+1} \|f\|_{C^{k+1}(\Delta_m)} n^{-(k+1)/2}. \end{aligned}$$

Then we consider the second term of the right-hand side of (18).

$$\begin{aligned} & \left\| \sum_{\|\alpha\|_1=k+1} (B_n - I)^{\lceil (k+1)/2 \rceil - 1} \left\{ \frac{\partial^\alpha f(\xi)}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ & \lesssim_{k+1} \|(B_n - I)^{\lceil (k+1)/2 \rceil - 1}\|_\infty \|f\|_{C^{k+1}(\Delta_m)} \sum_{\|\alpha\|_1=k+1} \sum_{\nu \in \bar{\Delta}_m} \left| \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha \right| b_{n,\nu}(\mathbf{q}) \\ & \lesssim_{k+1} \|(B_n - I)^{\lceil (k+1)/2 \rceil - 1}\|_\infty \|f\|_{C^{k+1}(\Delta_m)} \sum_{\|\alpha\|_1=k+1} \left\{ \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^{2\alpha} b_{n,\nu}(\mathbf{q}) \right\}^{1/2} \\ & \lesssim_{k+1,m} \|(B_n - I)^{\lceil (k+1)/2 \rceil - 1}\|_\infty \|f\|_{C^{k+1}(\Delta_m)} n^{-(k+1)/2}. \end{aligned}$$

Finally, we have

$$\begin{aligned} \|(B_n - I)^{\lceil (k+1)/2 \rceil} (f)(\mathbf{q})\|_\infty & \leq \sum_{\|\alpha\|_1=1}^k \left\| (B_n - I)^{\lceil (k+1)/2 \rceil - 1} \left\{ \frac{\partial^\alpha f(\mathbf{q})}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ & \quad + \left\| \sum_{\|\alpha\|_1=k+1} (B_n - I)^{\lceil (k+1)/2 \rceil - 1} \left\{ \frac{\partial^\alpha f(\xi)}{\alpha!} \sum_{\nu \in \bar{\Delta}_m} \left(\frac{\nu}{n} - \mathbf{q}\right)^\alpha b_{n,\nu}(\mathbf{q}) \right\} \right\|_\infty \\ & \lesssim_{k+1,m} \sum_{\|\alpha\|_1=1}^k \|(B_n - I)^{\lceil \|\alpha\|_1/2 \rceil - 1}\|_\infty \|f\|_{C^{k+1}(\Delta_m)} n^{-(k+1)/2} \\ & \quad + \|(B_n - I)^{\lceil (k+1)/2 \rceil - 1}\|_\infty \|f\|_{C^{k+1}(\Delta_m)} n^{-(k+1)/2} \\ & \lesssim_{k+1,m} \|f\|_{C^{k+1}(\Delta_m)} n^{-(k+1)/2}. \end{aligned}$$

The last inequality holds when $\|B_n - I\|_\infty$ is bounded by a constant independent of n . In fact, $\|(B_n - I)f\|_\infty = \sup_{\mathbf{q} \in \Delta_m} |(B_n - I)f(\mathbf{q})| = \sup_{\mathbf{q} \in \Delta_m} |\sum_{\nu \in \bar{\Delta}_m} \{f(\nu/n) - f(\mathbf{q})\} b_{n,\nu}(\mathbf{q})| \leq 2\|f\|_\infty$ and $\|B_n - I\|_\infty = \sup_{f \in C^k(\Delta_m), \|f\|_\infty \leq 1} \|(B_n - I)f\|_\infty \leq \sup_{f \in C^k(\Delta_m), \|f\|_\infty \leq 1} 2\|f\|_\infty \leq 2$. $\square$

Proof of Theorem 2. The first claim follows from Lemma 1 and the fact that $\max_{s \in [m]} \|g_s\|_{C^{2k}(\Delta_m)} \leq G$ and $\mathbb{E}_{X^n} \{D_{n,k}(g_s)(\mathbf{T}/n)\} = C_{n,k}(g_s)(\mathbf{q})$.

Additionally, we have

$$\begin{aligned} D_{n,k}(g_s)(\mathbf{q}) & = \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j B_n^j(g_s)(\mathbf{q}) \\ & = \sum_{j=0}^{k-1} \binom{k}{j+1} (-1)^j \{g_s(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1})\} \\ & = g_s(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1}), \end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{X^n} \left[\{D_{n,k}(g_s)(\mathbf{T}/n)\}^2 \right] &= \sum_{\boldsymbol{\nu} \in \tilde{\Delta}_m} \{D_{n,k}(g_s)(\boldsymbol{\nu}/n)\}^2 b_{n,\boldsymbol{\nu}}(\mathbf{q}) \\
&= B_n \left[\{D_{n,k}(g_s)\}^2 \right] (\mathbf{q}) \\
&= \{D_{n,k}(g_s)(\mathbf{q})\}^2 + \mathcal{O}_{k,m,G}(n^{-1}) \\
&= \{g_s(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1})\}^2 + \mathcal{O}_{k,m,G}(n^{-1}) \\
&= g_s^2(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{Var}_{X^n} \{D_{n,k}(g_s)(\mathbf{T}/n)\} &= \mathbb{E}_{X^n} \left[\{D_{n,k}(g_s)(\mathbf{T}/n)\}^2 \right] - [\mathbb{E}_{X^n} \{D_{n,k}(g_s)(\mathbf{T}/n)\}]^2 \\
&= g_s^2(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1}) - \{g_s(\mathbf{q}) + \mathcal{O}_{k,m,G}(n^{-1})\}^2 \\
&= \mathcal{O}_{k,m,G}(n^{-1}).
\end{aligned}$$

□

C Proof of Theorem 3

Proof of Theorem 3. By Theorem 2, it suffices to let $\tilde{P}_{X^n}(x = u_s | y = y^*) = D_{n,k}(g_s)(\mathbf{T}/n)$. Moreover, we have $\sum_{s=1}^m D_{n,k}(g_s)(\mathbf{T}/n) = 1$. □