Unbiased Test Error Estimation in the Poisson Means Problem via Coupled Bootstrap Techniques

Natalia L. Oliveira[1,2] Jing Lei[1] Ryan J. Tibshirani[1,2]

[1]Department of Statistics and Data Science, [2]Machine Learning Department
Carnegie Mellon University

Abstract

We propose a *coupled bootstrap* (CB) method for the test error of an arbitrary algorithm that estimates the mean in a Poisson sequence, often called the Poisson means problem. The idea behind our method is to generate two carefully-designed data vectors from the original data vector, by using synthetic binomial noise. One such vector acts as the training sample and the second acts as the test sample. To stabilize the test error estimate, we average this over multiple bootstrap B of the synthetic noise. A key property of the CB estimator is that it is unbiased for the test error in a Poisson problem where the original mean has been shrunken by a small factor, driven by the success probability p in the binomial noise. Further, in the limit as $B \to \infty$ and $p \to 0$, we show that the CB estimator recovers a known unbiased estimator for test error based on Hudson's lemma, under no assumptions on the given algorithm for estimating the mean (in particular, no smoothness assumptions). Our methodology applies to two central loss functions that can be sused to define test error: Poisson deviance and squared loss. Via a bias-variance decomposition, for each loss function, we analyze the effects of the binomial success probability and the number of bootstrap samples and on the accuracy of the estimator. We also investigate our method empirically across a variety of settings, using simulated as well as real data.

1 Introduction

We study the problem of estimating the test error of an algorithm in the Poisson many means problem, also called the Poisson compound decision problem. The importance of test error estimation in general rests on the fact that such estimates can be used in many dowstream applications, such as model assessment, selection, or tuning. To fix notation, given a data vector $Y = (Y_1, \dots Y_n) \in \mathbb{Z}_+^n$ (where we write $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ for the nonnegative integers) distributed according to

$$Y_i \sim \text{Pois}(\mu_i), \quad \text{independently}, \quad \text{for } i = 1, \dots, n, \tag{1}$$

we seek to estimate the mean vector $\mu = (\mu_1, \dots, \mu_n) \in \mathbb{R}_+^n$ (where we use $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ for the set of nonnegative real numbers). Let $g : \mathbb{Z}_+^n \to \mathbb{R}_+^n$ be a measurable function that estimates μ from the data Y, so that we can write $\hat{\mu} = g(Y)$. We will often refer to g as an algorithm, in the context of estimating μ in the Poisson many means problem.

To evaluate the performance of g, we can use various metrics. One class of metrics evaluate what we call *test error*, based on a loss function $L : \mathbb{Z}_+^n \times \mathbb{R}_+^n \to \mathbb{R}$,

$$\text{Err}(g) = \mathbb{E}[L(\tilde{Y}, g(Y))], \quad \text{where } \tilde{Y} \text{ is drawn from (1), independently of } Y, \tag{2}$$

which measures how well g tracks an independent copy $\tilde{Y}$ of the data. A second class of metrics evaluate what we call *risk*, again based on a loss function L,

$$\text{Risk}(g) = \mathbb{E}[L(\mu, g(Y))], \tag{3}$$

which measures how well g tracks the mean $\mu = \mathbb{E}[Y]$ of the data. Admittedly, many authors use the terms "test error" and "risk" interchangeably, but in this paper we are careful to use terminology that distinguishes the two, for reasons that we will become apparent in the next subsection.

1.1 Test error versus risk

In the classical normal means problem, where instead of (1) we observe $Y_i \sim N(\mu_i, \sigma^2)$ independently, it is straightforward to show that under a squared loss L, the test error (2) and risk (3) differ only by the noise level σ^2. In the Poisson means problem, there is no direct analogy for typical loss functions of interest, and the difference between (2) and (3) will generally depend on μ. This means that an estimator of one metric (test error or risk) does not as easily translate into an estimator of the other, since μ is of course unknown, and the primary estimand of interest.

Thankfully, as we show here, when L is a Bregman divergence the difference between test error and risk does not depend on g. A Bregman divergence is a loss function of the form $L(a,b) = D_\phi(a,b)$, where

$$D_\phi(a,b) = \phi(a) - \phi(b) - \langle \nabla\phi(b), a - b \rangle, \tag{4}$$

for a convex, differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$, where here an subsequently we use $\langle u, v \rangle = u^\mathsf{T} v$ for vectors u, v. In this case it is straightforward to see that

$$\begin{aligned}
\mathrm{Err}(g) - \mathrm{Risk}(g) &= \mathbb{E}[D_\phi(\tilde{Y}, g(Y))] - \mathbb{E}[D_\phi(\mu, g(Y))] \\
&= \mathbb{E}[\phi(\tilde{Y})] - \mathbb{E}[\phi(g(Y))] - \mathbb{E}[\langle \nabla\phi(g(Y)), \tilde{Y} - g(Y) \rangle] \\
&\quad - \phi(\mu) + \mathbb{E}[\phi(g(Y))] + \mathbb{E}[\langle \nabla\phi(g(Y)), \mu - g(Y) \rangle] \\
&= \mathbb{E}[\phi(Y)] - \phi(\mu),
\end{aligned} \tag{5}$$

where the cancellation of terms in the third line holds because $Y, \tilde{Y}$ are i.i.d., and thus $\mathbb{E}[\langle \nabla\phi(g(Y)), \tilde{Y} \rangle] = \langle \mathbb{E}[\nabla\phi(g(Y))], \mu \rangle$. Observe that (5) is the gap in Jensen's inequality. Therefore it is always nonnegative, and $\mathrm{Err}(g) \geq \mathrm{Risk}(g)$.

In this paper, we will focus on estimating the test error (2) in the Poisson means problem (1), for two special instances of a Bregman divergence: squared loss and Poisson deviance, as will be discussed in the next subsection. Since $\mathrm{Err}(g) - \mathrm{Risk}(g) = \mathbb{E}[\phi(Y)] - \phi(\mu)$ depends on μ, it will not be the case that we can automatically translate an estimator of the test error of g into an estimator of its risk. However, we can still unbiasedly estimate the *difference* in risk between two models g and h, as discussed next.

Model comparisons. The gap (5) does not depend on g. Thus for a comparison between two algorithms g and h, we always have (provided we use Bregman divergence to define the test error and risk metrics):

$$\mathrm{Err}(g) - \mathrm{Err}(h) = \mathrm{Risk}(g) - \mathrm{Risk}(h),$$

To be clear, this means that if $\widehat{\mathrm{Err}}(g)$ is an unbiased estimator of $\mathrm{Err}(g)$ for any g, just as we will produce in this paper, then

$$\widehat{\mathrm{Err}}(g) - \widehat{\mathrm{Err}}(h) \quad \text{is unbiased for} \quad \mathrm{Risk}(g) - \mathrm{Risk}(h), \quad \text{for any } g, h.$$

As such, we can still use the tools developed in this paper to perform model comparisons, or more broadly, model tuning (where g_s is indexed by a tuning parameter $s \in S$, and we select s to minimize an unbiased estimate of test error, or equivalently, risk).

Fixed-X Poisson regression. A special case of our problem setting to keep in mind is *fixed-X* Poisson regression. Here we view $Y \in \mathbb{R}^n$ as a response vector and we have an associated feature matrix $X \in \mathbb{R}^{n \times p}$. The algorithm g typically performs a kind of Poisson regression of Y on X. As long as we consider X to be fixed (nonrandom), we can still interpret this as a problem of the form (1), with $\mu = \mu(X)$. In this setting, the test error metric (2) translates to what is called fixed-X prediction error, where we evaluate predictions at the same feature vectors (rows of X), but against new responses (elements of $\tilde{Y}$).

While fixed-X analyses are more typical in classical statistics, the *random-X* perspective is great interest in modern prediction problems. Here the feature vectors at which we make predictions are random, giving rise to random-X prediction error as the metric of concern. Estimating random-X prediction error is *not* in general equivalent to estimating fixed-X prediction error and the two can behave quite differently (see, e.g., Rosset and Tibshirani (2020) for an extended discussion). The random-X perspective eludes the framework of the current paper, but is an important topic for future work.

1.2 Squared loss versus Poisson deviance

When $\phi(x) = \|x\|_2^2$, it is easy to check that

$$D_\phi(a, b) = \|a - b\|_2^2,$$

which is the squared loss. Meanwhile, when $\phi(x) = 2\sum_{i=1}^n x_i(\log x_i - 1)$, it follows that

$$D_\phi(a, b) = 2\sum_{i=1}^n \bigg(a_i \log \frac{a_i}{b_i} + b_i - a_i\bigg),$$

which is known as *Poisson deviance*. We will take these to be the two loss functions of primary interest in our work. Accordingly, we introduce the notation for test error under squared loss and Poisson deviance:

$$\mathrm{Err}^{\mathrm{sqr}}(g) = \mathbb{E}\|\tilde{Y} - g(Y)\|_2^2, \tag{6}$$

$$\mathrm{Err}^{\mathrm{dev}}(g) = 2\mathbb{E}\bigg[\sum_{i=1}^n \bigg(\tilde{Y}_i \log \frac{\tilde{Y}_i}{g_i(Y)} + g_i(Y) - \tilde{Y}_i\bigg)\bigg]. \tag{7}$$

Squared loss is a standard choice in many estimation and prediction problems and does not really need further motivation. Poisson deviance can be motivated from different perspectives; one nice perspective is that, if we parametrize $g_i(Y) = \exp(\theta_i)$ for $i = 1, \dots, n$, then fitting g to minimize Poisson deviance on the given data is equivalent to maximum likelihood in the Poisson model,

$$\underset{g}{\text{minimize}}\ 2\bigg[\sum_{i=1}^n \bigg(Y_i \log \frac{Y_i}{g_i(Y)} + g_i(Y) - Y_i\bigg)\bigg] \iff \underset{\theta}{\text{minimize}} \sum_{i=1}^n \Big(-Y_i\theta_i + \exp(\theta_i)\Big).$$

In the same vein, evaluating g by Poisson deviance on $\tilde{Y}$ is equivalent to evaluating g by Poisson likelihood on an independent copy of the training sample.

In our view, squared loss and Poisson deviance are each important loss functions, and are each deserving of study. This is only strengthened by the fact that they can have very different behaviors in certain problem settings. As a simple example, suppose $n = 1$, and we have two scenarios: in the first $\tilde{Y} = 1$ and $g(Y) = 2$, while in the second $\tilde{Y} = 500$ and $g(Y) = 501$. The squared loss in each scenario is 1. However, the Poisson deviance in first scenario is ≈ 0.307, and in the second scenario it is ≈ 0.001. The difference here is driven by the fact that in the Poisson model the variance scales with the mean. Hence according to Poisson deviance (equivalent to Poisson likelihood), a prediction of 502 when the predictand is 501 is not nearly as bad as a prediction of 2 when the predictand is 1.

In Sections 5 and 6, we will present and discuss several examples that expose differences in the behavior of squared loss and Poisson deviance in different settings. That said, our primary focus is on estimating test error defined with respect to these loss functions, and not on comparing them. A comprehensive analysis of their differences is beyond the scope of the current paper.

1.3 Hudson's lemma

A fundamental result in this area is *Hudson's lemma*, due to Hudson (1978). Hudson actually derived two identities, one each for continuous and discrete exponential families. These can be viewed as extensions of Stein's celebrated identity (Stein, 1981) for the Gaussian family.[1] For concreteness, we state Hudson's result for the Poisson case.

Lemma 1 (Hudson 1978)**.** *Let $Y_i \sim \mathrm{Pois}(\mu_i)$, independently, for $i = 1, \dots, n$. Let $g : \mathbb{Z}_+^n \to \mathbb{R}^n$ be such that $\mathbb{E}|g_i(Y)| < \infty$, $i = 1, \dots, n$. Then, denoting by $e_i \in \mathbb{R}^n$ the vector whose i^{th} entry is 1, with all others 0,*

$$\mu_i \mathbb{E}[g_i(Y)] = \mathbb{E}[Y_i g_i(Y - e_i)], \quad i = 1, \dots, n, \tag{8}$$

where by convention we set $g_i(-1) = 0$, $i = 1, \dots, n$.

[1] Stein's work was actually completed as a technical report in 1973, and was a motivation for Hudson's work, even though the publication dates of their papers do not reflect this. According to Hudson, Stein already knew of the result in (8).

Compared to Stein's identity, which requires that g is weakly differentiable, Hudson's identity (8) holds without any smoothness assumptions on g (of course, even formulating precisely what smoothness would mean over a discrete domain like $\mathbb{Z}_n^+$ would be tricky, but the lack of assumptions needed for Lemma 1 are remarkable nonetheless). Hudson's main interest was in developing inadmissibility results for estimators of the location parameter in an exponential family distribution. The identities he established were used as tools in his analysis, which parallels Stein's use of his own identity in Stein (1981).

Moreover, analogous to what can be done with Stein's lemma, Hudson's lemma can be used to derived unbiased estimators for various risk metrics in exponential families. An important contribution in this area is Eldar (2009), and further contributions (along with a comprehensive summary of available tools and results from the literature) are given in Deledalle (2017).

1.4 Unbiased estimation

Our focus in this paper is slightly unique, since we consider test error (2) as the primary target and not risk (3), as considered by Eldar (2009); Deledalle (2017), and most other authors in the literature. Nonetheless, the estimators developed by these authors have natural analogues for test error. In fact, the story is for test error is simpler, and an unbiased estimator can be obtained for any Bregman divergence loss function.

To see this, we first recall a general decomposition of test error for Bregman divergence losses known as *Efron's optimism theorem*, due to Efron (1975, 1986, 2004): this shows that for any Bregman divergence D_ϕ in (4) and any algorithm g, this difference in test error and training error satisfies

$$\begin{aligned}\mathbb{E}[D_\phi(\tilde{Y}, g(Y))] - \mathbb{E}[D_\phi(Y, g(Y))] &= \mathbb{E}[\phi(\tilde{Y})] - \mathbb{E}[\phi(g(Y))] - \mathbb{E}[\langle \nabla\phi(g(Y)), \tilde{Y} - g(Y)\rangle] \\ &\quad - \mathbb{E}[\phi(Y)] + \mathbb{E}[\phi(g(Y))] + \mathbb{E}[\langle \nabla\phi(g(Y)), Y - g(Y)\rangle] \\ &= \mathbb{E}[\langle \nabla\phi(g(Y)), Y\rangle] - \mathbb{E}[\langle \nabla\phi(g(Y)), \mu\rangle]. \end{aligned} \tag{9}$$

The second line follows from the fact that $\mathbb{E}[\phi(\tilde{Y})] = \mathbb{E}[\phi(Y)]$.[2] Simply rewriting the above, we see that if we are able to construct an unbiased estimator $V(g)$ of $\mathbb{E}[\langle \nabla\phi(g(Y)), \mu\rangle]$ then

$$D_\phi(Y, g(Y)) + \langle \nabla\phi(g(Y)), Y\rangle - V(g)$$

will be an unbiased estimator for $\mathbb{E}[D_\phi(\tilde{Y}, g(Y))]$. In the Poisson case, Hudson's identity (8) precisely gives the unbiased estimator $V(g)$ that we require, which leads to the following result.

Proposition 1. *Let $Y_i \sim \mathrm{Pois}(\mu_i)$, independently, for $i = 1, \ldots, n$. Let $g : \mathbb{Z}_+^n \to \mathbb{R}^n$ be any algorithm and D_ϕ be any Bregman divergence loss function (indexed by a convex, differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$) such that $\mathbb{E}|\phi(g(Y))| < \infty$ and $\mathbb{E}|\nabla_i \phi(g(Y))| < \infty$, $i = 1, \ldots, n$. Then*

$$\mathrm{UE}(g) = D_\phi(Y, g(Y)) + \langle \nabla\phi(g(Y)), Y\rangle - \langle \nabla\phi(g_-(Y)), Y\rangle \tag{10}$$

is unbiased for $\mathrm{Err}(g) = \mathbb{E}[D_\phi(\tilde{Y}, g_i(Y))]$, where we abbreviate $g_-(Y) = (g_1(Y - e_1), \ldots, g_n(Y - e_n))$, and as usual, $\tilde{Y}$ denotes an independent copy of Y.

As a consequence, we have the following unbiased estimators for squared loss and Poisson deviance:

$$\mathrm{UE}^{\mathrm{sqr}}(Y) = \|Y\|_2^2 + \|g(Y)\|_2^2 - 2\langle g_-(Y), Y\rangle, \tag{11}$$

$$\mathrm{UE}^{\mathrm{dev}}(Y) = 2\sum_{i=1}^n \Big(Y_i \log Y_i - Y_i \log g_i(Y - e_i) + g_i(Y) - Y_i \Big). \tag{12}$$

These are altogether highly similar to the unbiased risk estimators in Eldar (2009); Deledalle (2017), and to be clear, we do not consider (11), (12) to be major (or even original) contributions of our work. That said, we have not yet seen the general unbiased estimator for Bregman divergence (10) noted in the literature, thus we believe it may be useful to record it (along with the observation that estimation of test error can be easier than estimation of risk).

[2] This exposes the reason why the analogous decomposition for risk can be more complex: when we replace $\tilde{Y}$ with μ in the calculation that led to (9), we are left with an extra term $\phi(\mu) - \mathbb{E}[\phi(Y)]$ that does not cancel and must be estimated.

The estimators in (11), (12) have a clear strength: they are unbiased *for any algorithm* g. This is a strong property; recall that by comparison, in the Gaussian model, the analogous estimator is Stein's unbiased risk estimator (SURE), which requires g to be weakly differentiable. The estimators in (11), (12) also have a clear downside: they require the algorithm g to be run $n+1$ times, once to obtain the original fit $g(Y)$, and then n more times to obtain $g_-(Y)$, which recall has entries $g_i(Y-e_i)$, $i=1,\dots,n$. Thus we can liken (11), (12) to leave-one-out cross-validation, in terms of computational cost.

This draws a clear line of motivation to the main contribution of our paper: in what follows, we develop an unbiased estimator of test error, for any Bregman divergence loss function D_ϕ and any algorithm g, using a carefully-crafted parametric bootstrap scheme. The computational cost (number of runs of g) here is tied to a user-controlled parameter B, the number of bootstrap samples. In general, increasing B decreases the variance of the estimator, but any choice of $B \geq 1$ yields an estimator that is unbiased for the test error in a *mean-shrunken* Poisson problem, with mean $(1-p)\mu$, where $p > 0$ is another user-controlled parameter.

1.5 Summary of contributions

The following gives a summary of our main contributions and an outline for this paper.

- In Section 2, we introduce the coupled bootstrap (CB) estimator, and prove that it is unbiased for the test error in a mean-shrunken Poisson problem.
- In Section 3, we analyze the behavior of the CB estimator as $B \to \infty$ and $p \to 0$, and prove that the limiting CB estimator is exactly the unbiased estimator (10) from Hudson's lemma.
- In Section 4, we study the bias and variance of the CB estimator and quantify how they depend on B, p and other problem parameters.
- In Section 5, we compare the CB and the unbiased estimator on various simulated data sets, and find that the performance of the CB estimator is favorable, especially when the algorithm g is unstable.
- In Section 6, we examine the use of the CB estimator for model tuning—selecting from a family g_s, $s \in S$ of algorithms—in two applications: image denoising and density estimation. We find that using Poisson deviance (to define the test error metric) consistently delivers more regularized models than using squared loss.
- In Section 7, we conclude with a brief discussion and ideas for future work.

1.6 Related work

Estimating risk and test error is of central importance in statistics and machine learning. In the random-X prediction setting (which recall does not fit in the framework of our work) the most ubiquitous estimator is arguably cross-validation, which itself carries a long line of literature. We do not describe this literature here, but highlight Bates et al. (2021) as a nice recent paper that carefully reexamines this classic estimator, and also provides a nice overview of literature on cross-validation.

In the fixed-X prediction setting—or in general, parametric many means problems—there has also been a long history of work in statistics, with Akaike (1973); Mallows (1973); Efron (1975); Stein (1981); Efron (1986) marking early important contributions. This has been particularly well-studied in the Gaussian means problem, and in this area, we draw attention to Breiman (1992); Ye (1998); Efron (2004), and particularly to Oliveira et al. (2021), as motivation for our current work in the Poisson means problem. These papers use auxiliary noise—they inject synthetic (Gaussian) noise into the data at hand—in order to estimate the risk or fixed-X prediction error of an arbitrary algorithm g. Our previous work, Oliveira et al. (2021), proposes a coupled bootstrap (CB) scheme for doing so that has a simple, intuitive target of estimation for any auxiliary noise level. In particular, for any auxiliary noise level $\alpha > 0$ (a user-controlled parameter), the CB method produces an unbiased estimator for the risk in a Gaussian means problem that has an inflated noise variance $(1+\alpha)\sigma^2$ (where σ^2 denotes the original noise level). The current paper builds off this idea, and develops a coupled boostrap scheme in the Poisson model that enjoys analogous properties.

Relative to the Gaussian case, risk and test error estimation in the Poisson means model has been less well-studied. However, there has still certainly been important and influential work in the area. This includes

Hudson (1978), as already described in the introduction, and also Shen et al. (2004); Eldar (2009); Deledalle (2017). Meanwhile, the literature on *mean estimation*—the role played by what we are calling the algorithm g—in the Poisson many means problem is vast. Quite a lot of work on this topic has been done in the signal processing community, where it is often called Poisson denoising; see, e.g., Harmany et al. (2009); Luisier et al. (2010); Raginsky et al. (2010); Harmany et al. (2012); Salmon et al. (2014); Cao and Xie (2016). Therefore we believe that the techniques we develop for estimating test error in the Poisson means model should have widespread practical applications, in signal processing and elsewhere.

Lastly, we mention a concurrent, related line of work on auxiliary randomization approaches that allow for rigorous post-selection inference in parametric many means models, including the Poisson means model. We highlight Leiner et al. (2021); Neufeld et al. (2023) as two nice recent papers in the area. In particular, a core piece of the auxiliary randomization procedure in our work was directly inspired by the former paper, and their use of binomial auxiliary noise in the Poisson model.

2 Coupled boostrap estimator

In this section, we introduce the CB estimator in the Poisson means model, and investigate some of its basic properties.

2.1 Proposed estimator

The following is a simple but key "three-point" formula for expected Bregman divergence loss from Oliveira et al. (2021) that will drive our main proposal in this paper.

Proposition 2. *Let $U, V, W \in \mathbb{R}^n$ be independent random vectors. For any g, and Bregman divergence D_ϕ,*

$$\mathbb{E}[D_\phi(V, g(U))] - \mathbb{E}[D_\phi(W, g(U))] = \mathbb{E}[\phi(V)] - \mathbb{E}[\phi(W)] + \langle \mathbb{E}[\nabla\phi(U)], \mathbb{E}[W] - \mathbb{E}[V] \rangle, \tag{13}$$

assuming all expectations exist and are finite. In particular, if U, V are i.i.d. and $\mathbb{E}[U] = \mathbb{E}[W]$, then

$$\mathbb{E}[D_\phi(V, g(U))] = \mathbb{E}[D_\phi(W, g(U))] + \mathbb{E}[\phi(U)] - \mathbb{E}[\phi(W)]. \tag{14}$$

Proof. The first statement (13) follows from the definition of Bregman divergence (4), and the independence of U, V, W. The second result (14) follows from the first, by noting that if U, V are i.i.d. and $\mathbb{E}[U] = \mathbb{E}[W]$ then $\mathbb{E}[V] = \mathbb{E}[W]$, thus the last term on the right-hand side in (13) is zero, and the first term is $\mathbb{E}[\phi(U)]$. □

While simple to state and prove, the results in Proposition 2 are useful observations. To map them onto to the problem of estimating of test error in the Poisson model, consider the following. Given a Poisson data vector Y from (1), suppose that we can generate a pair of vectors $(U, W) = (Y^*, Y^\dagger)$ that are independent of each other and have the same mean. Then (14) says that

$$D_\phi(Y^\dagger, g(Y^*)) + \phi(Y^*) - \phi(Y^\dagger) \quad \text{is unbiased for} \quad \mathbb{E}[D_\phi(\tilde{Y}^*, g(Y^*))], \tag{15}$$

where $\tilde{Y}^*$ is an independent copy of Y^*. In other words, the above constructs an unbiased esitmator for the test error in a problem in which the original data vector was Y^*, rather than Y. Thus if Y^* *was "close" in distribution to* Y, then this estimator would be meaningful. (Ideally, we would like Y^* to be be identical in distribution to Y, but that will not be generically possible without knowledge of μ.)

What remains is a precise scheme in the Poisson setting to generate the pair $(Y^*, Y^\dagger)$ from Y such that $Y^*, Y^\dagger$ are independent, $\mathbb{E}[Y^*] = \mathbb{E}[Y^\dagger]$, and Y^*, Y are "close" in distribution. The next lemma does the trick and fulfills these three properties precisely. It was brought to our attention by Leiner et al. (2021) who used it in a distinct but generally related post-selection inference context. For completeness, we provide a proof in Appendix A.1. Here and henceforth we use the following abbreviations: we write $Y \sim \text{Pois}(\mu)$ to mean that we draw $Y_i \sim \text{Pois}(\mu_i)$, independently, for $i = 1, \ldots, n$, and similarly $Z \sim \text{Binom}(N, p)$ to mean that we draw $Z_i \sim \text{Binom}(N_i, p)$, independently, for $i = 1, \ldots, n$.

Lemma 2. *Given $Y \sim \text{Pois}(\mu)$, fix any $0 < p < 1$ and let $\omega \,|\, Y \sim \text{Binom}(Y, p)$. Then, defining $Y^* = Y - \omega$ and $Y^\dagger = (1-p)/p \cdot \omega$, it holds that:*

(i) $Y^*, Y^\dagger$ *are independent;*

(ii) $\mathbb{E}[Y^*] = \mathbb{E}[Y^\dagger]$*; and*

(iii) $Y^* \sim \text{Pois}((1-p)\mu)$.

Lemma 2, combined with the observation in (15), forms the basis for the CB test error estimator in the Poisson many means problem. To stabilize the estimator, we can simply repeat the draws of binomial noise from Lemma 2 over independent repetitions $b = 1, \dots, B$. To be concrete, this leads to the following method: we first generate samples according to

$$\begin{aligned} \omega^b \,|\, Y &\sim \text{Binom}(Y, p), \quad \text{independently}, \quad \text{for } b = 1, \dots, B, \\ Y^{*b} = Y - \omega^b, \quad &, Y^{\dagger b} = \frac{1-p}{p}\omega^b, \quad \text{for } b = 1, \dots, B, \end{aligned} \tag{16}$$

for an arbitrary binomial success probability $0 < p < 1$, and a number of bootstrap draws $B \geq 1$; then we define the *coupled bootstrap* (CB) estimator, for test error under Bregman divergence loss D_ϕ, by:

$$\text{CB}_p(g) = \frac{1}{B} \sum_{b=1}^{B} \Big(D_\phi(Y^{\dagger b}, g(Y^{*b})) + \phi(Y^{*b}) - \phi(Y^{\dagger b}) \Big). \tag{17}$$

We can view each Y^{*b} as a synthetic training set for g, and each $Y^{\dagger b}$ as a synthetic test set. The correction term $\phi(Y^{*b}) - \phi(Y^{\dagger b})$ accounts for the fact that $Y^{*b}, Y^{\dagger b}$ do not have the same distribution (though recall they do have the same mean, by construction).

For the two loss functions of primary interest, squared loss and Poisson deviance, the CB estimator in (17) becomes, respectively:

$$\text{CB}_p^{\text{sqr}} = \frac{1}{B} \sum_{b=1}^{B} \Big(\|Y^{\dagger b} - g(Y^{*b})\|_2^2 + \|Y^{*b}\|_2^2 - \|Y^{\dagger b}\|_2^2) \Big), \tag{18}$$

$$\text{CB}_p^{\text{dev}} = \frac{2}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \Big(Y_i^{*b} \log Y_i^{*b} - Y_i^{\dagger b} \log g_i(Y^{*b}) + g_i(Y^{*b}) - Y_i^{\dagger b} \Big). \tag{19}$$

Interlude: special care with deviance estimators. We take a brief but practically important detour to note that special care must be taken with test error estimators with respect to Poisson deviance loss. In this case, each of the unbiased (12) and the coupled bootstrap (19) estimators can diverge if the coordinate functions of g can output zero. For the unbiased estimator this occurs when $Y_i \neq 0$ and $g_i(Y - e_i) = 0$; for the coupled bootstrap estimator this occurs when $Y_i^{\dagger b} \neq 0$ and $g_i(Y^{*b}) = 0$. As a safety mechanism, we can simply pad the output of g so that zero is never in the range of its coordinate functions: say, we can define a modified algorithm

$$\tilde{g}_i(y) = g_i(y) 1\{g_i(y) \neq 0\} + c 1\{g_i(y) = 0\}, \quad i = 1, \dots, n,$$

for a small constant $c > 0$. This is reasonable because even the population Poisson deviance (7) can itself diverge when the coordinate functions of g can output zero. With a modified rule like the one above, we may ask how frequently the padding is actually in effect in the computation of the estimators (12) and (19). We study this in Appendix A.2 and show that, in a sense, it is typically in effect less frequently in the CB estimator (19) than in the unbiased one (12).

2.2 Unbiasedness for mean-shrunken target

The next result is immediate from Lemma 2 and (15).

Corollary 1. *Let* $Y \sim \text{Pois}(\mu)$*. Let* $g : \mathbb{Z}_+^n \to \mathbb{R}^n$ *be any algorithm, let* D_ϕ *be any Bregman divergence loss function, and let* $0 < p < 1$ *and* $B \geq 1$ *be arbitrary. Then the CB estimator* $\text{CB}_p(g)$ *in* (17) *is unbiased for* $\text{Err}_p(g)$ *(assuming all terms in* (17) *have finite expectations), where* $\text{Err}_p(g)$ *is the test error of* g *with respect to a mean-shrunken Poisson problem:*

$$\text{Err}_p(g) = \mathbb{E}[D_\phi(\tilde{Y}_p, g(Y_p))], \quad \textit{where } Y_p, \tilde{Y}_p \sim \text{Pois}((1-p)\mu), \textit{ and } Y_p, \tilde{Y}_p \textit{ are independent.} \tag{20}$$

The strength of Corollary 1 rests on the fact that the estimand in (17) of $\mathrm{CB}_p(g)$ for any choice of $p > 0$ is highly intuitive: it is $\mathrm{Err}_p(g)$ in (20), which is the test error that we would encounter in a slightly harder version of our original problem, where the mean μ has been replaced by $(1-p)\mu$.

Why is this important? It means that we do *not* have to send $p \to 0$ in order to be able to interpret the estimand of the CB estimator, and thus justify its use. Any nonzero (noninfinitesimal) p will still result in a target that has a clear, intuitive meaning. This is good news for the CB estimator, because when p is away from zero, we can generally choose a reasonably small number of bootstrap draws B in order to stabilize the variance of the estimator, which presents a computational advantage over the unbiased estimator in (10). We will learn more about the behavior of the CB estimator, as we vary p and B, in Sections 4 and 5 (where we formally analyze the bias and variance, and carry out empirical comparisons, respectively).

2.3 Smoothness of mean-shrunken target

Now that we have shown that $\mathrm{CB}_p(g)$ is unbiased for $\mathrm{Err}_p(g)$, it is natural to ask whether $\mathrm{Err}_p(g)$ will be close to $\mathrm{Err}(g)$ for small p. Our next result gives a partial answer by proving that if g satisfies some mild moment conditions, then the map $p \mapsto \mathrm{Err}_p(g)$ will be continuous (and in fact, it can be continuously differentiable, depending on the number of moments assumed) in an interval containing $p = 0$. Later on, in Section 4, we will derive results that give a more quantitative sense of how close $\mathrm{Err}_p(g)$ can be to $\mathrm{Err}(g)$.

Proposition 3. *For $0 \leq p < 1$, let $\mathrm{Err}_p(g)$ be as defined in (20). If for some integer $k \geq 0$,*

$$\mathbb{E}\big[D_\phi(\tilde{Y}, g(Y))\langle \tilde{Y} + Y, 1_n\rangle^m\big] < \infty, \quad m = 0, \ldots, k,$$

where in the above above $Y, \tilde{Y} \sim \mathrm{Pois}(\mu)$ are independent, and $1_n \in \mathbb{R}^n$ denotes the vector of all 1s, then the map $p \mapsto \mathrm{Err}_p(g)$ has k continuous derivatives on $[0, 1)$.

The proof of this result is not difficult but a bit technical and deferred to Appendix A.3. We remark that when $k = 0$, the assumption in Proposition 3 is simply $\mathrm{Err}(g) = \mathbb{E}[D_\phi(\tilde{Y}, g(Y))] < \infty$ (i.e., the original test error is finite), which is extremely weak, and even in this case we get that $\mathrm{Err}_p(g) \to \mathrm{Err}(g)$ as $p \to 0$.

3 Noiseless limit

In this section, we consider the *infinite-bootstrap* version of the CB estimator, $\mathrm{CB}_p^\infty(g) = \lim_{B\to\infty} \mathrm{CB}_p(g)$. By the law of large numbers, this estimator is equivalent to taking the expectation over the binomial noise,

$$\mathrm{CB}_p^\infty(g) = \mathbb{E}[\mathrm{CB}_p(g) \,|\, Y] = \mathbb{E}\bigg[D_\phi\bigg(\frac{1-p}{p}\omega, g(Y-\omega)\bigg) + \phi(Y-\omega) - \phi\bigg(\frac{1-p}{p}\omega\bigg)\bigg], \tag{21}$$

where $\omega \,|\, Y \sim \mathrm{Binom}(Y, p)$, as in (16).

The next result considers the *noiseless limit* of the infinite-bootstrap version of the CB estimator (21), where $p \to 0$. Its proof is deferred until Appendix B.

Theorem 1. *Let $Y \sim \mathrm{Pois}(\mu)$. Let $g : \mathbb{Z}_+^n \to \mathbb{R}^n$ be any algorithm, let D_ϕ be any Bregman divergence loss function, and assume that $|\phi_i(g(Y))| < \infty$, $|\nabla_i \phi(g(Y))| < \infty$, and $|\nabla_i \phi(g(Y - e_i))| < \infty$ almost surely, for each $i = 1, \ldots, n$. Then*

$$\lim_{p\to 0} \mathrm{CB}_p^\infty(g) = \mathrm{UE}(g), \quad \textit{almost surely}, \tag{22}$$

where $\mathrm{UE}(g)$ is the unbiased estimator defined in (10). Thus as a consequence, the noiseless limit of $\mathrm{CB}_p^\infty(g)$ is unbiased for $\mathrm{Err}(g)$.

That the limiting CB estimator (17) recovers the unbiased estimator (10) based on Hudson's lemma, as $B \to \infty$ and $p \to 0$, is certainly an encouraging property for the former. We recall that in the Gaussian many means problem, the analogous result was derived in Oliveira et al. (2021): there, the CB estimator recovers the unbiased estimator based on Stein's lemma, in the noiseless limit. However, this Gaussian result requires g to be weakly differentiable (which is the condition required for Stein's unbiased estimator to be valid in the first place). In the current Poisson many means problem, note that Theorem 1 requires no such restrictions on g (and indeed, recall, the unbiased estimator does not either, from Proposition 1).

Figure 1 illustrates this difference via a simple simulation; see the figure caption for details.

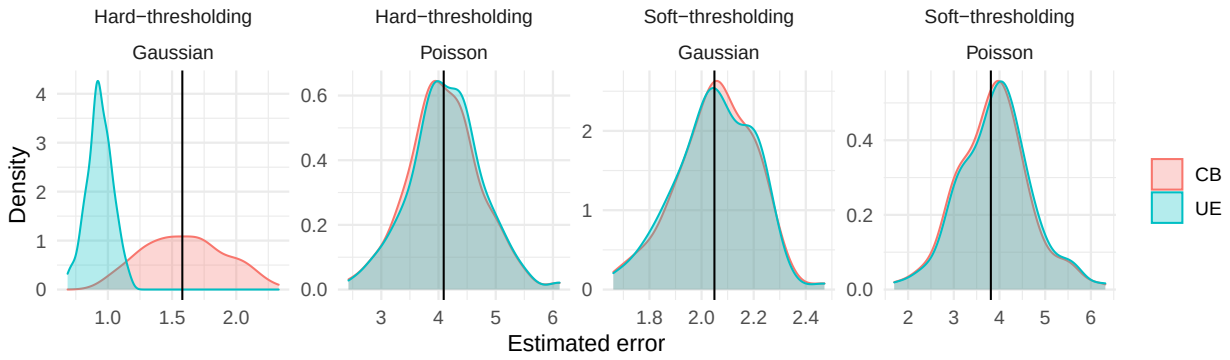


Figure 1: Density of the CB and unbiased estimators of test error in Gaussian and Poisson settings, where the CB estimators effectively take large B and a small amount of auxiliary noise. This is based on a simulation with $n = 30$, where we generate Gaussian or Poisson data with constant mean, and consider two estimators: soft-thresholding and hard-thresholding (the latter violating weak differentiability). The black vertical line in each panel marks the true test error. For small auxiliary noise in the Gaussian hard-thresholding setting, it is clear that CB and the unbiased estimator are separated (and the latter is far from unbiased).

4 Bias and variance

In this section, we study a bias-variance decomposition of the estimator $\mathrm{CB}_p(g)$ in (17), when targeting the *original* error $\mathrm{Err}(g)$. We will consider an arbitrary Bregman divergence loss (used to define $\mathrm{Err}(g)$), and use the decomposition

$$\mathbb{E}[\mathrm{CB}_p(g) - \mathrm{Err}(g)]^2 = \underbrace{\big[\mathrm{Err}_p(g) - \mathrm{Err}(g)\big]^2}_{\mathrm{Bias}^2(\mathrm{CB}_p(g))} + \underbrace{\mathbb{E}\big[\mathrm{Var}(\mathrm{CB}_p(g) \,|\, Y)\big]}_{\mathrm{RVar}(\mathrm{CB}_p(g))} + \underbrace{\mathrm{Var}\big(\mathbb{E}[\mathrm{CB}_p(g) \,|\, Y]\big)}_{\mathrm{IVar}(\mathrm{CB}_p(g))}. \tag{23}$$

This is the usual bias-variance decomposition of squared error loss, where we have used $\mathbb{E}[\mathrm{CB}_p(g)] = \mathrm{Err}_p(g)$ in the bias term, and we have further expanded the usual variance term (using the law of total variance) into two components which we call the *reducible* and *irreducible* variance, respectively, as in Oliveira et al. (2021). We note that as the number of bootstrap draws B grows, the reducible variance shrinks, but the irreducible variance does not; the latter does not depend on B at all, and in fact, it can be viewed as the variance of the infinite-bootstrap version of the estimator, $\mathrm{CB}_p^\infty(g) = \mathbb{E}[\mathrm{CB}_p(g) \,|\, Y]$.

In what follows, we will analyze each of the three terms in (23) to understand their behavior as functions of p and B, with a focus on small p and large B. As usual, we assume throughout that $Y \sim \mathrm{Pois}(\mu)$, where $Y_p \sim \mathrm{Pois}((1-p)\mu)$ for $p \geq 0$, and we denote by $\tilde{Y}, \tilde{Y}_p$ independent copies of Y, Y_p, respectively. Lastly, D_ϕ represents an arbitrary Bregman divergence loss.

4.1 Bias

First we give an exact expression for the bias, $\mathrm{Bias}(\mathrm{CB}_p(g)) = \mathrm{Err}_p(g) - \mathrm{Err}(g)$, and an upper bound on its magnitude for small p, under an assumption of monotone variance. The proof is given in Appendix C.1.

Proposition 4. *Assume that* $\mathbb{E}[D_\phi(\tilde{Y}_p, g_p(Y))\langle Y_p + \tilde{Y}_p, 1_n\rangle] < \infty$. *Then for all* $p \in [0, 1)$,

$$\mathrm{Err}_p(g) - \mathrm{Err}(g) = -\sqrt{2\sum_{i=1}^n \mu_i} \int_0^p \frac{1}{\sqrt{1-t}} \mathrm{Cor}\Big(D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big) \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}_t, g(Y_t))\big]}\, dt. \tag{24}$$

Further, if $\mathrm{Var}[D_\phi(\tilde{Y}_p, g(Y_p))]$ *is decreasing in* p *on* $[0, 1/2]$, *then for any* p *in this range,*

$$|\mathrm{Err}_p(g) - \mathrm{Err}(g)| \leq \frac{5p}{3} \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}, g(Y))\big] \sum_{i=1}^n \mu_i} \tag{25}$$

We remark that the assumption of decreasing variance of the loss is fairly natural (because the variance of each component of Y_p decreases monotonically to 0 as p increases to 1). We can also drop this condition, and replace the variance term in the bound (25) by $\sup_{t\in[0,p)} \mathrm{Var}[D_\phi(\tilde{Y}_t, g(Y_t))]$.

4.2 Reducible variance

Next we bound the reducible variance, $\mathrm{RVar}(\mathrm{CB}_p(g))$. We focus on the dependence on p and B, for small p and large B. The notation $O(\cdot)$ is to be interpreted in this regime (small p, large B), and hides factors that may depend on the mean μ, which may in turn depend on the dimensionality n. The proof is given in Appendix C.2.

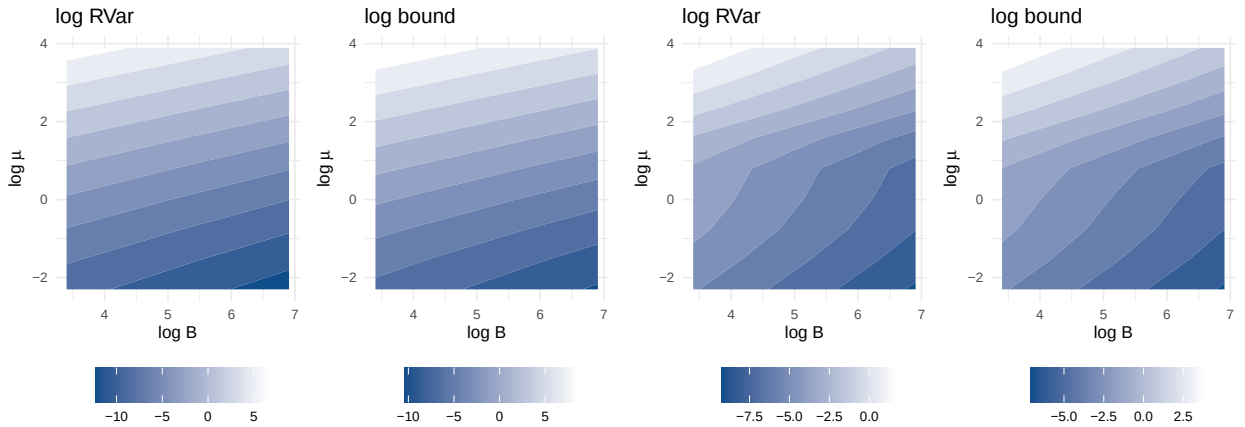
Proposition 5. *Assume the variables $f(Y_p)$, $f^2(Y_p)$, $f(Y_p)\langle Y_p, 1_n\rangle$, $f^2(Y_p)\langle Y_p, 1_n\rangle$ all have finite L^1 norm, uniformly bounded over all functions $f \in \mathcal{F}$ and all $p \in [0, q)$, for some $q > 0$, where*

$$\mathcal{F} = \big\{y \mapsto D_\phi(y, g(y))\big\} \cup \big\{y \mapsto \nabla_i \phi(g(y)) : i = 1, \ldots, n\big\}.$$

Then for all $p \in [0, q)$,

$$\mathrm{RVar}(\mathrm{CB}_p(g)) \le \frac{2}{B}\mathrm{Var}\Big[D_\phi(Y, g(Y))) + \langle Y, \nabla\phi(g(Y))\rangle\Big] + \frac{2}{Bp}\sum_{i=1}^n \mu_i \mathbb{E}\big[\nabla_i\phi(g(Y_p))^2\big] + \frac{2}{B}\sum_{i=1}^n \mu_i^2 \mathrm{Var}\big[\nabla_i\phi(g(Y_p))\big] + O\Big(\frac{p}{B}\Big). \quad (26)$$

A simple simulation, whose results are presented in Figure 2, shows that the reducible variance bound (26) appears to have the right dependence on μ and B. See the figure caption for details.



(a) Squared loss

(b) Deviance loss

Figure 2: Comparison of the true reducible variance (approximated by Monte Carlo) and the bound given in (26), for squared and deviance loss, in a simulation with $n = 100$ and $p = 0.1$. The data vector Y has Poisson entries, and μ denotes the common mean of each component; we use a simple linear shrinkage estimator g. We can see that the behavior for varying B, μ looks qualitatively similar across the true reducible variance heatmap and the bound heatmap, for each loss function.

4.3 Irreducible variance

Last we analyze the irreducible variance, $\mathrm{IVar}(\mathrm{CB}_p(g))$. The proof is given in Appendix C.3.

Proposition 6. *Assume that* $\mathbb{E}[D_\phi^2(Y, g(Y))] < \infty$ *and* $\mathbb{E}[\langle \nabla(g(Y)), Y\rangle^2] < \infty$. *Define* Φ_g *to have component functions*

$$\Phi_{g,i}(y) = \sup_{0 \le z \le y} \big|\nabla_i \phi(g(y))\big|, \quad i = 1, \ldots, n.$$

(Here when we write $0 \le z \le y$*, all inequalities are to be interpreted componentwise.) Then,*

$$\lim_{p \to 0} \mathrm{IVar}(\mathrm{CB}_p(g)) \le 2\mathrm{Var}\Big[D_\phi(Y, g(Y)) + \langle \nabla\phi(g(Y)), Y\rangle\Big] + 2\mathbb{E}\big[\langle \Phi_g(Y), Y\rangle^2\big]. \tag{27}$$

4.4 Discussion of bias and variance results

We discuss interpretation of the results above. The bias bound (25) decreases linearly with p, which suggests that we should take p to be as small as possible in order to decrease the bias. The irreducible variance bound (27) provides no resistance to this idea, as it has a stable noiseless limit, as we send $p \to 0$. The behavior of the reducible variance bound (26), however, is more intricate. The second term on the right-hand side in (26) diverges as $p \to 0$, but this can be offset by sending $B \to \infty$.

How large do we need to take B? Altogether, there are really only two quantities on the right-hand side in (26) that B needs to balance out, which are the second and third terms. First, let us normalize the target error by the number of samples, because this would be the natural scale of concern, in general (our original definition of $\mathrm{Err}(g)$ in (6) or (7) is a sum, rather than an average, over samples). We can see from (23) that rescaling each of $\mathrm{Err}(g)$ and $\mathrm{CB}_p(g)$ by $1/n$ multiplies all terms in the error decomposition—bias, reducible variance, and irreducible variance—by a factor of $1/n^2$. Now, ignoring constants, the (squared) bias bound (25) and the second and third terms in the reducible variance bound (26) are, after multiplying by $1/n^2$:

$$\frac{p^2}{n^2}\|\mu\|_1 \quad \text{and} \quad \frac{1}{n^2 B p}\|\mu\|_1 + \frac{1}{n^2 B}\|\mu\|_2^2,$$

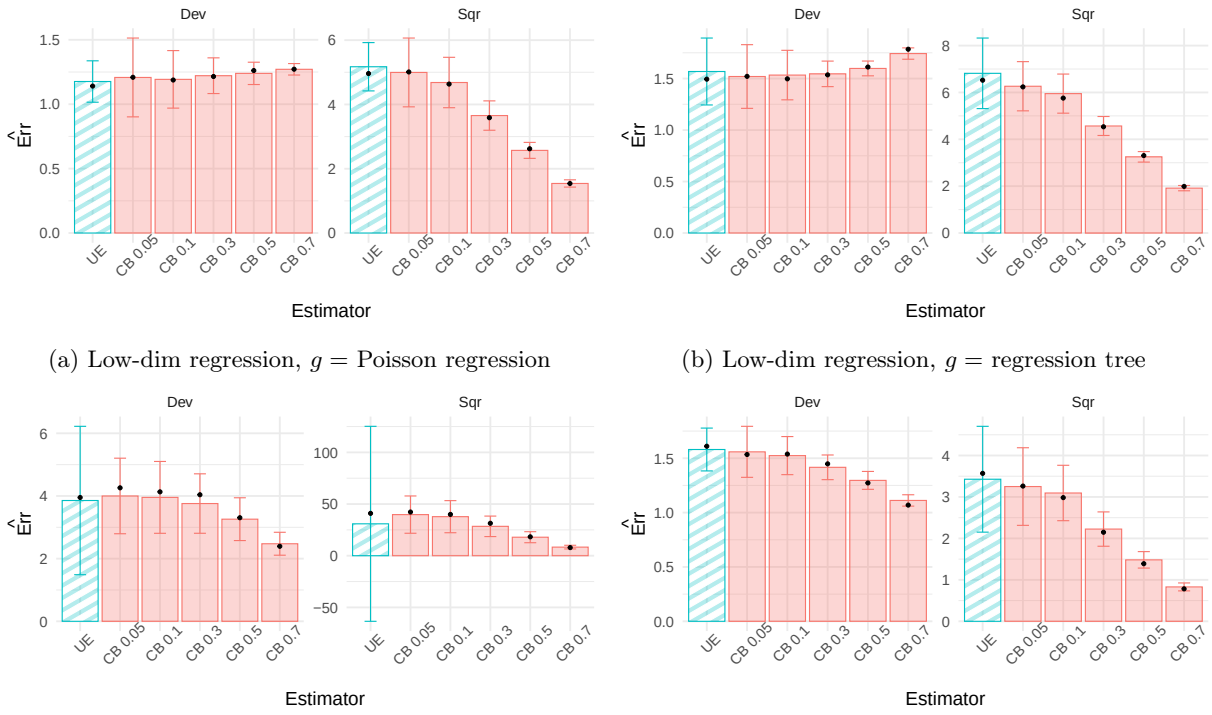
respectively, where recall, we use $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n_+$ for mean vector. As we can see, increasing the total signal energy $\|\mu\|_1$ adversely affects the control we have over the bias and reducible variance. In a moderate signal regime, where $\|\mu\|_1/n$ is moderate or small, the rough orders for the bias and the reducible variance in the above display would be small, even for only modest values of p and B. However, in a large signal regime, where $\|\mu\|_1/n$ is large (possibly increasing as the sample size n grows), we may need to take p to be small to offset this (if we want the bias to be held small), which requires us to take B large enough to dominate $\|\mu\|_1/(n^2 p)$ or $\|\mu\|_2^2/n^2$ (depending on which is larger) in the reducible variance bound.

In practice, for any given problem at hand, we would generally recommend choosing p to be small, such as $p = 0.05$ or $p = 0.1$, but not tiny. This choice is made in favor of keeping the variance under control (for a reasonable number of bootstrap samples B), at the potential expense of incurring a nontrivial bias in the CB estimator. However, this brings us back to a primary feature of the CB estimator—recall, for any $p > 0$, it is unbiased for $\mathrm{Err}_p(g)$. This represents a shift in focus, where we now consider estimating error in a problem setting where the mean has been shrunk from μ to $(1-p)\mu$, which is intuitively a conservative bet and often a reasonable undertaking even for moderately small but not infinitesimal values of p.

5 Simulated experiments

In this section, we run and analyze two sets of simulations. The first, presented in Section 5.1, compares the unbiased estimator (UE) in (10) and the CB estimator in (17), across four settings. Each setting is defined by a different data model and algorithm g, and we examine the performance of CB versus UE in estimating the true error, as we vary the binomial noise parameter p, for a fixed sample size n. We find that CB performs favorably overall: it delivers similar error estimates to UE for small values of p, and importantly, it can have much smaller variance than UE when g is unstable.

The second set of simulations, presented in Section 5.2, focuses on just one setting in which UE generally behaves favorably. The motivation here is to compare the variability of CB and UE after stratifying the two to have roughly equal computational cost—which is accomplished by sampling summands in (11) or (12). In this simulation, we fix the binomial noise parameter p, and vary the sample size n and signal size μ. We find that CB has lower variability unless the signal size μ is very large.



(a) Low-dim regression, g = Poisson regression

(b) Low-dim regression, g = regression tree

(c) High-dim regression, g = 5-fold CV lasso

(d) Denoising, g = EB 1-step estimator

Figure 3: Comparison of CB and UE across different data models, algorithms, and loss functions.

5.1 CB versus UE, varying p

Here we compare the CB and UE estimators, for squared and deviance loss functions: see (18), (19) for CB and (11), (12) for UE. Throughout, we set $n = 100$, and use $B = 100$ bootstrap samples for CB. We consider the following combinations of different data models for Y, and algorithms g:

- *Low-dimensional regression.* We set $p = 10$, draw features $X_i \sim N(\theta, I_p)$, independently, $i = 1, \ldots, n$, where each $\theta_j = 3$ and I_p denotes the $p \times p$ identity matrix; then we draw responses $Y_i \sim \text{Pois}(X_i^T \beta)$, independently, $i = 1, \ldots, n$, where each $\beta_j = 0.05$. This corresponds to a signal-to-noise ratio (SNR) of approximately 2. We examine two choices for g, a Poisson regression and a regression tree.

- *High-dimensional regression.* We set $p = 200$, and use a similar setup to the above, except with features $X_i \sim N(0, \sigma^2 I_p)$, where $\sigma^2 = 1.5$, and responses $Y_i \sim \text{Pois}(X_i^T \beta)$, where each $\beta_j = 0.13$. This was done to maintain an SNR of roughly 2. In this setting, we take g to be a lasso Poisson regression, with the tuning parameter λ chosen by 5-fold cross-validation (CV).

- *Denoising.* We draw $Y_i \sim \text{Pois}(\mu_i)$, independently, $i = 1, \ldots, n$, where $\mu_i = 10$ for $i \leq 10$ and $\mu_i = 0.5$ for $i > 10$. In this setting, we take g to be a 1-step improvement on an empirical Bayes (EB) estimator as described in Brown et al. (2013), with tuning parameter fixed at $h = 0.85$.

In each setting, we perform 100 repetitions (i.e., we draw the data vector $Y \in \mathbb{R}^n$ 100 times from the specified data model); but we note that in the regression settings, the features are drawn once and fixed throughout. We consider a range of noise levels p for the CB estimator: 0.05, 0.1, 0.3, 0.5, and 0.7. All error metrics and error estimators, here and throughout all empirical examples, are scaled by $1/n$.

The results are displayed in Figure 3, with each panel (a)–(d) displaying a different combination of data model and algorithm g. In each panel, the average test error estimate is displayed for each method (CB or UE), as well as standard errors measured over the 100 repetitions. Furthermore, the black points denote the true estimands (computed via Monte Carlo): $\text{Err}(g)$ for UE, and $\text{Err}_p(g)$ for CB. As expected, all estimators

are seen to be roughly unbiased for their targets: $\mathrm{Err}(g)$ or $\mathrm{Err}_p(g)$. Interestingly, we also see that for squared loss, the target $\mathrm{Err}_p(g)$ clearly decreases as p grows, but for deviance loss, the behavior of $\mathrm{Err}_p(g)$ tends to be more robust to growing p (and can increase or decrease, depending on the setting).

In panel (a), which is the low-dimensional regression setting, with Poisson regression as the algorithm g, we can see that UE has a noticeably lower standard error than CB at the lowest binomial noise level of $p = 0.05$, particularly for deviance loss. This is the only setting in which this happens. In all others, the CB estimator at the lowest noise level has either comparable or smaller variability than UE. In fact, in panel (c), which is the high-dimensional regression setting, with CV-tuned lasso as the algorithm g, we see that UE has a dramatically higher standard error than CB at any level of noise p. The algorithm g is inherently unstable here, because CV (operating in high-dimensions, and at a moderate SNR) can choose very different tuning parameter values across different data instances. Despite this, CB is able to deliver estimates of reasonably low variance, since it averages across draws of auxiliary binomial noise, which acts as a method of smoothing (like bagging). We note that the analogous phenomenon also occurs in the Gaussian setting, as observed by Oliveira et al. (2021).

5.2 CB versus UE, sampling summands

The unbiased estimator in (10) requires $n + 1$ runs of the algorithm g, making it computationally expensive for large sample sizes. In contrast, the number of runs of g required by the CB estimator in (17) is B, which is a user-chosen parameter (recall that any choice of B results in an unbiased estimator $\mathrm{CB}_p(g)$ for $\mathrm{Err}_p(g)$, whereas larger B reduces the variance of the estimator). In the last subsection, we fixed $n = B = 100$. In this one, we consider larger much sample sizes, with n ranging from 10^3 to 10^5. We maintain $B = 100$, but we equate computational costs between UE and CB by sampling $m = 100$ summands uniformly at random (and without replacement) from (11) or (12), and then scaling up the resulting sum by n/m. We denote the estimator resulting from this "sampling summands" approach by $\mathrm{UE}_{\mathrm{ss}}(g)$, which is unbiased for $\mathrm{Err}(g)$.

For the data model, we draw $Y_i \sim \mathrm{Pois}(\mu)$, independently, $i = 1, \ldots, n$, where μ ranges from 0.5 to 30. For the algorithm, we use a simple linear shrinkage estimator: $g(y) = 0.8y + 0.2\bar{y} + 0.011\{\bar{y} = 0\}$. We note that this choice is generally favorable to UE, and more unstable algorithms g would only create more variability for UE relative to CB, and thus look more favorable to CB, as observed in the last subsection. For the binomial noise parameter, we set $p = \min\{0.1, \sum_{i=1}^n \mu_i / \sum_{i=1}^n \mu_i^2\}$, which roughly balances the leading terms in the reducible variance upper bound (26).

The results are displayed in Figure 4. For deviance loss, the results are overall quite favorable for CB: it has a lower variance than $\mathrm{UE}_{\mathrm{ss}}$ (darker shade of blue) in all but the top left corner, which corresponds to small n and large μ. In fact, the variability of CB is quite similar (across all n, μ) to that of UE for deviance loss, even though the former is considerably cheaper ($B = 100$ runs of the algorithm g, versus n runs). For squared loss, there is more of a clear tradeoff: for large values of μ (i.e., roughly $\log \mu > 2$, or $\mu > 7.38$), we see that CB has greater variability than $\mathrm{UE}_{\mathrm{ss}}$; for moderate values of μ (roughly $\log \mu$ between 0 and 2, or μ between 1 and 7.38), CB has comparable variability for small n and smaller variability for large n; while for small values of μ (roughly $\log \mu < 0$, or $\mu < 1$), CB has smaller variability than $\mathrm{UE}_{\mathrm{ss}}$.

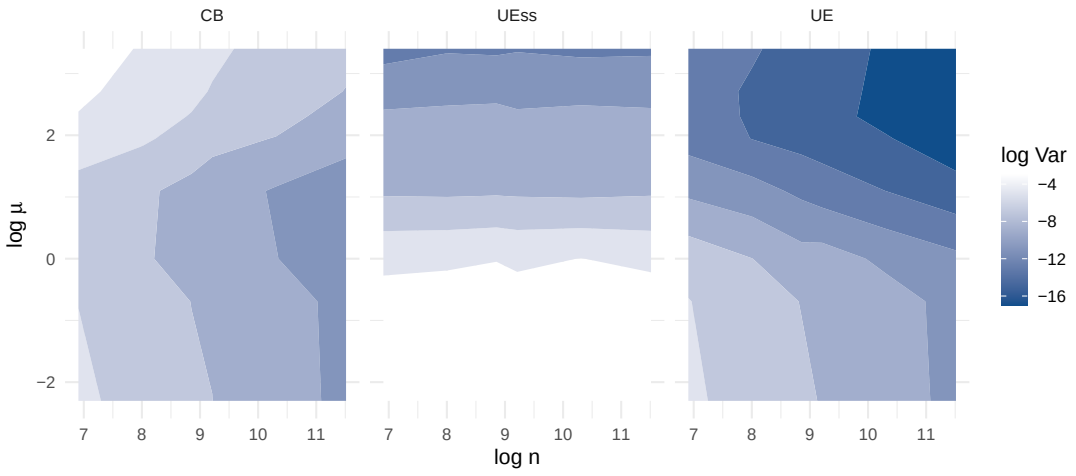
6 Applications

6.1 Image denoising

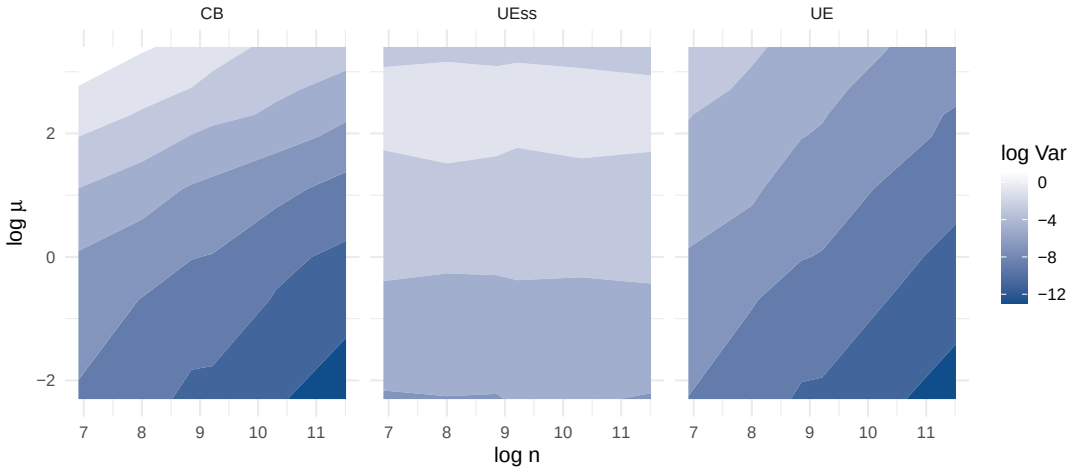
We consider the following Poisson image denoising framework from Harmany et al. (2012) (motivated by the study of Poisson noise or shot noise in areas such as microscopy and astrophotography). We observe data $Y_i \sim \mathrm{Pois}(f_i^*)$, independently, $i = 1, \ldots, n$, where $f^* \in \mathbb{R}_+^n$ is an unknown signal of interest, which we assume has the structure of an $N \times N$ image, where $n = \sqrt{N}$. We consider an estimator $\hat{f} = g(Y)$ for f^* given by solving the optimization problem:

$$\underset{f \geq 0}{\text{minimize}} \sum_{i=1}^n \big(- Y_i \log(f_i + \rho) + f_i + \rho \big) + \tau \sum_{i \sim j} |f_i - f_j|, \tag{28}$$

where ρ is a small positive constant to avoid the singularity $f = 0$, and we write $i \sim j$ to indicate that indices i, j are adjacent to each other in the ordering determined by the underlying image. This estimator is a form



(a) Deviance loss



(b) Squared loss

Figure 4: Variability of CB, UE_{ss}, and UE as functions of n and μ, for a simple linear shrinkage estimator.

of total variation (TV) regularized Poisson image denoising; we note that the loss term is equivalent (up to constants) to the Poisson deviance between f and Y, and the penalty term encourages the estimated image $\hat{f}$ to be piecewise constant, with the tuning parameter $\tau \geq 0$ determining the strength of regularization.

As an example, we consider the well-known synthetic phantom image f^*, of resolution 128×128 (so that $n = 16384$). We use CB to estimate the test error of $\hat{f}$, the Poisson image denoising estimator defined in (28), over a range of values of the tuning parameter τ. We consider both squared (18) and deviance (19) loss, and set $B = 50$ and $p = 0.1$. We did not consider the unbiased estimator in (11) or (12) in this experiment, due to its prohibitive computational cost (it requires $n = 16384$ refits of the TV denoising estimator (28)).

Figure 5 displays the CB error curve and true error curve (approximated by Monte Carlo), as functions of τ, with separate panels for squared and deviance loss. There are two points worth noting. First, despite the gap between the CB and true test error curves (unsurprising, because $p = 0.1$), their curvature is similar; in particular, the value of τ minimizing the CB curve is close to the value minimizing test error. This is the case for both squared and deviance loss, and it shows that the CB estimator can be useful for model tuning, even when p is not small. Second, the value of τ minimizing the CB curve is larger for deviance loss than it is for squared loss, marked by the dotted lines in each panel. This translates into a greater degree of regularization, as can be seen clearly in Figure 6, which plots the denoised estimates themselves at the CB-optimal values of τ, for squared and deviance loss.

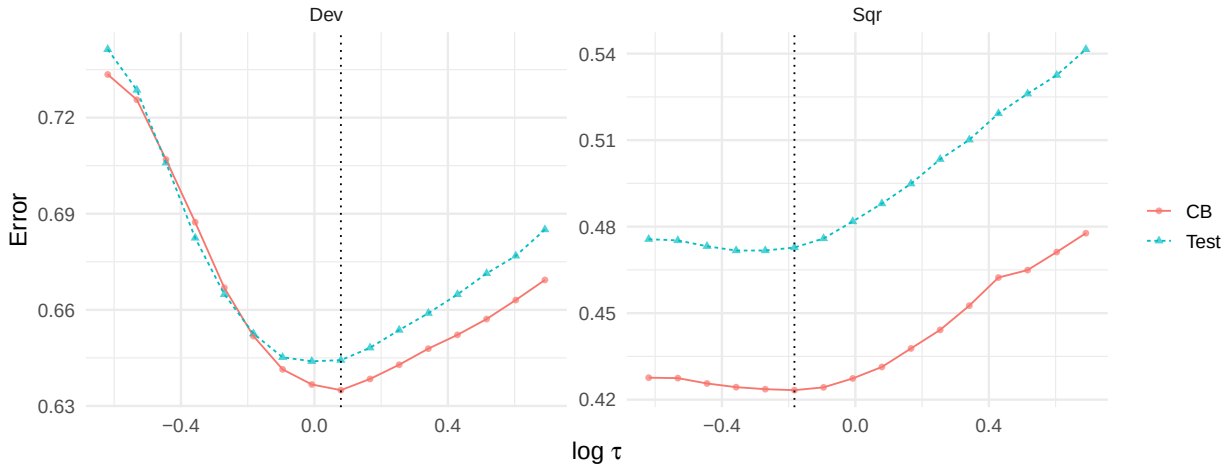


Figure 5: Comparison of CB and true test error curves, as functions of the tuning parameter τ, for a Poisson image denoising estimator.

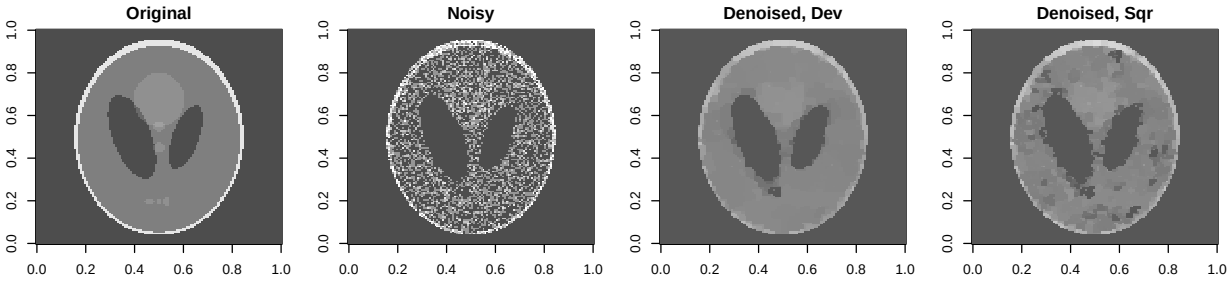


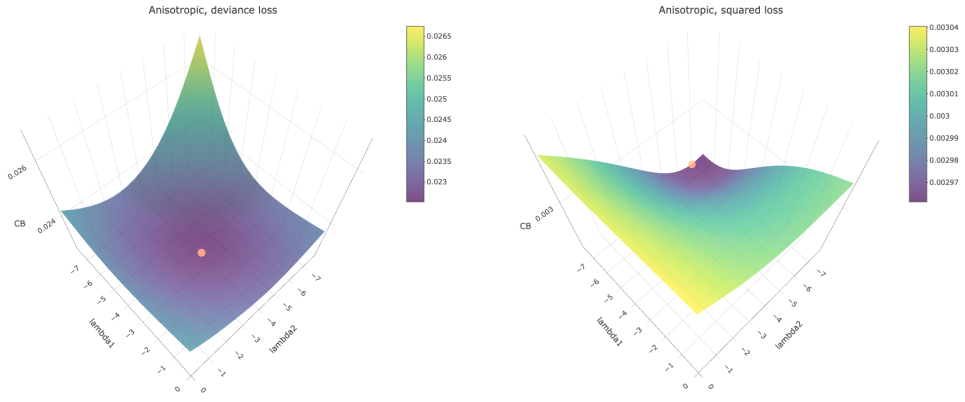
Figure 6: Original, noisy, and denoised estimates at the CB-optimal value of τ, for squared and deviance loss.

6.2 Density estimation

We study density estimation, which can be turned into a Poisson regression via Lindsey's method (Lindsey, 1974; Lindsey and Mersch, 1992; Efron and Tibshirani, 1996). The basic idea is to discretize the domain into bins, and model the count in each bin as a Poisson random variable, with the mean parameter constrained or regularized to be smoothly varying across the bins. We can then estimate the mean parameter by (regularized) maximum likelihood, which in turn gives a discretized density estimate.

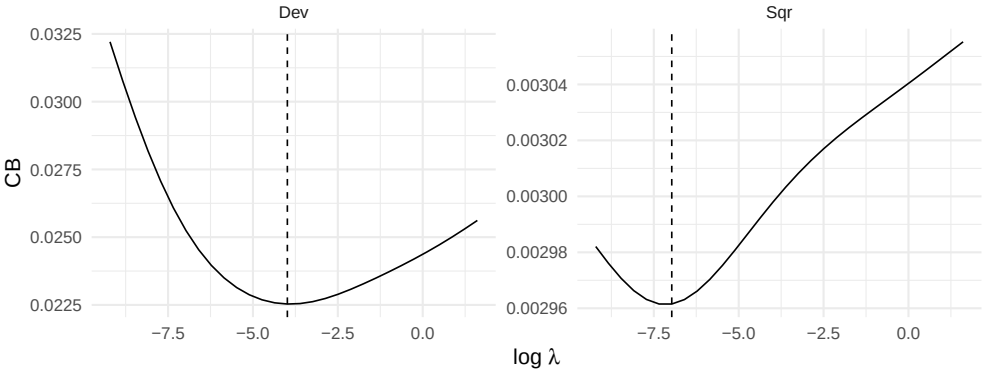
In particular, we consider an example from Phillips et al. (2006) on the distribution of *Bradypus variegatus*, a lowland species of sloth found across Central and South America. Each data point consists of latitude and longitude pair, representing a site where a sloth was seen, and the data set contains 116 total sightings. To form a 2d density estimate, we apply Lindsey's method, with 200 equally-spaced bins along the latitude and longitude axes, and we use a P-spline to model the Poisson mean parameter. P-splines were first proposed by Eilers and Marx (1996); details for the 2d case can be found in Eilers and Marx (2003); Eilers et al. (2006); Eilers and Marx (2021). We use a 2d cubic B-spline parametrization for the mean function with 30 knots in each dimension, and we use a penalty on the sum of squared second-order differences across adjacent B-spline parameters along each dimension. Moreover, we consider two versions of this penalty: an *anisotropic* version, which decouples the regularization strength along each dimension, and has two tuning parameters $\lambda_1, \lambda_2 \geq 0$; and an *isotropic* version, which ties together the regularization strength over the dimensions, and has a single tuning parameter $\lambda \geq 0$.

Figure 7 shows the results of using the CB method, with $B = 100$ and $p = 0.1$, to estimate both squared and deviance loss, across a range of tuning parameter values. For either the anisotropic or isotropic penalty, it is clear that minimizing CB-estimated deviance loss leads to larger tuning parameter values—and hence more regularized density estimates—than minimizing CB-estimated squared loss. As we can see in Figure 8, this



(a) Anisotropic penalty, deviance loss

(b) Anisotropic penalty, squared loss



(c) Isotropic penalty

Figure 7: CB curves for anisotropic and isotropic penalties as a function of the tuning parameter(s).

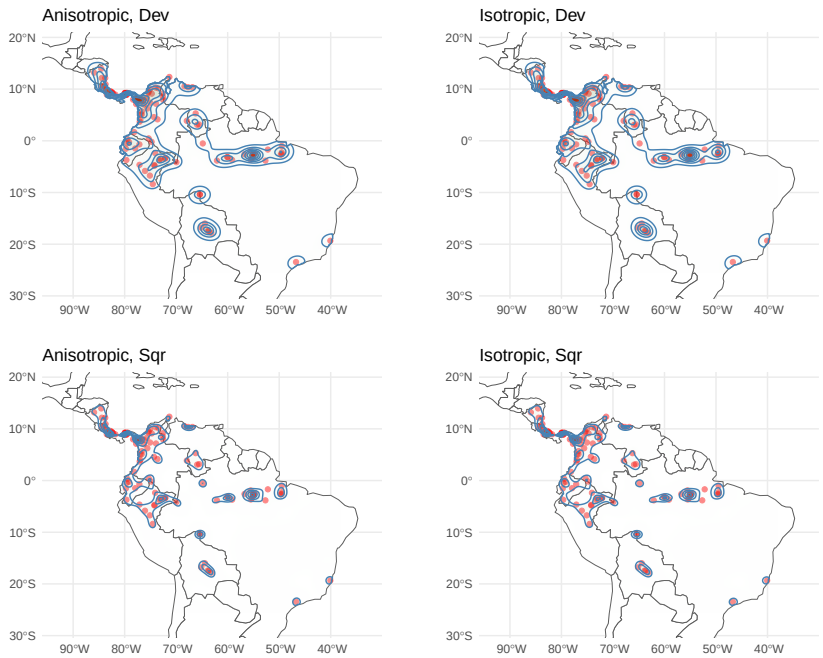


Figure 8: Density estimates at the CB-optimized values of the tuning parameters for anisotropic and isotropic penalties, and squared and deviance loss.

leads to more plausible looking density estimates (top row). The estimates obtained optimizing CB-estimated squared loss (bottom row) appear too concentrated around the observations themselves.

7 Discussion

We proposed and analyzed a coupled bootstrap (CB) method for test error estimation in the Poisson means problem, with a focus on squared and Poisson deviance loss functions. The CB estimator, for any choice of the binomial noise parameter $p > 0$, is unbiased for an intuitive target: $\mathrm{Err}_p(g)$, the test error of the given algorithm g, when the mean vector in the Poisson model has been shrunk from μ to $(1-p)\mu$. Importantly, this unbiasedness requires no assumptions on g whatsoever. Furthermore, we proved that in the noiseless limit $p \to 0$, the CB estimator (with infinite bootstrap iterations) reduces to the natural unbiased estimator (UE) for test error that comes from an application of Hudson's lemma. However, CB has two key advantages over UE. First, it requires running the algorithm g in question B times (which is a user-controlled parameter in CB), versus $n+1$ times (which comes directly from the form of UE). Second, as we show in our experiments, CB can often have smaller variance than UE, particularly when the underlying algorithm g is unstable.

We finish by emphasizing that it would be interesting to extend the CB framework to other data models, beyond Gaussian, as in Oliveira et al. (2021), and Poisson, as in the current paper. To explain what would be required for this, it may be helpful to first recap the general developments in Section 2. Given any random vector Y, suppose that we can generate a pair $(Y^*, Y^\dagger)$ such that:

(i) $Y^*, Y^\dagger$ are independent; and

(ii) $\mathbb{E}[Y^*] = \mathbb{E}[Y^\dagger]$.

Then letting $\tilde{Y}^*$ denote an independent copy of Y^*, Proposition 2 implies (as stated in (15), which we copy here for convenience):

$$D_\phi(Y^\dagger, g(Y^*)) + \phi(Y^*) - \phi(Y^\dagger) \quad \text{is unbiased for} \quad \mathbb{E}[D_\phi(\tilde{Y}^*, g(Y^*))],$$

for any Bregman divergence D_ϕ which serves as our loss function. Therefore, under properties (i) and (ii) we can estimate error as measured by an arbitrary Bregman divergence, unbiasedly—granted, the error here is defined when the training and test distributions are given by that of Y^*, which is different from our original data distribution. This means that there is actually an implicit third property that we need in order for us to want to use the estimator in the above display:

(iii) the law of Y^* is "close enough" to that of Y that $\mathbb{E}[D_\phi(\tilde{Y}^*, g(Y^*))]$ is an "interesting" proxy target.

This is less explicit than either (i) or (ii) but it is just as important. To be clear, the properties (i), (ii), and (iii) are already met by the existing Gaussian and Poisson constructions. Moreover, for any given distribution of Y, if we can fulfill (i), (ii), and (iii), then we can build a corresponding CB estimator for the (proxy) test error by averaging the above construction over multiple bootstrap draws, as in (17).

Towards satisfying properties (i) and (ii), the recent paper of Neufeld et al. (2023) provides a number of constructions which serve a related but distinct purpose, in a selective inference context. From some initial random vector Y, they seek to create a pair $(Y^{(1)}, Y^{(2)})$ which are independent, and satisfy $Y^{(1)} + Y^{(2)} = Y$. Fortunately, by simple rescaling, one can check that their constructions (from their Table 2) can be adapted to satisfy (i) and (iii) for the gamma, exponential, binomial, multinomial, and negative binomial families of distributions. Meanwhile, property (iii) can be argued on a case-by-case basis. As an example, consider $n = 1$ (only for simplicity, the same idea can be applied coordinatewise in the multivariate case), and assume $Y \sim \mathrm{Exp}(\lambda)$, exponentially distributed with rate $\lambda > 0$. Then for arbitrary $\epsilon \in (0,1)$, we can define

$$\begin{aligned} Z &\sim \mathrm{Beta}(\epsilon, 1-\epsilon), \\ Y^* &= \frac{Z}{\epsilon} \cdot Y, \\ Y^\dagger &= \frac{1-Z}{1-\epsilon} \cdot Y, \end{aligned}$$

where $\text{Beta}(\alpha, \beta)$ denotes the beta distribution with shapes $\alpha, \beta > 0$. From Neufeld et al. (2023), we know that $Y^*, Y^\dagger$ are independent, with $Y^* \sim \text{Gam}(\epsilon, \epsilon\lambda)$ and $Y^\dagger \sim \text{Gam}(1-\epsilon, (1-\epsilon)\lambda)$, where $\text{Gam}(\alpha, \beta)$ is the gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$. Thus, we can see that $\mathbb{E}[Y^*] = \mathbb{E}[Y^\dagger]$, and so (i) and (ii) are clearly satisfied. Furthermore, the distribution $\text{Gam}(\epsilon, \epsilon\lambda)$ of Y^* is indeed similar to that $\text{Exp}(\lambda)$ of Y, with the latter approaching the former as $\epsilon \to 1$, which confirms our property (iii).

Given the success we have seen for the CB method in the Gaussian and Poisson settings, we feel these and other extensions are worth exploring, along of course with theory and experiments to support their use as potentially core tools for error and risk estimation in denoising and fixed-X regression problems.

Acknowledgements

We thank Aaditya Ramdas and Boyan Duan for inspiring discussions. NLO was supported by an Amazon Fellowship.

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281, 1973.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? arXiv: 2104.00673, 2021.

Leo Breiman. The little bootstrap and other methobds for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

Lawrence D. Brown, Eitan Greenshtein, and Ya'acov Ritov. The Poisson compound decision problem revisited. *Journal of the American Statistical Association*, 108(502):741–749, 2013.

Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2016.

Charles-Alban Deledalle. Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Electronic Journal of Statistics*, 11(2):3141–3164, 2017.

Bradley Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3(6):1189–1242, 1975.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Bradley Efron and Robert Tibshirani. Using specially designed exponential families for density estimation. *Annals of Statistics*, 24(6):2431–2461, 1996.

Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11 (2):89–121, 1996.

Paul H. C. Eilers and Brian D. Marx. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66:159–174, 2003.

Paul H. C. Eilers and Brian D. Marx. *Practical Smoothing: The Joys of P-splines*. Cambridge University Press, 2021.

Paul H. C. Eilers, Iain D. Currie, and Maria Durbán. Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76, 2006.

Yonina C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

Zachary T. Harmany, Roummel F. Marcia, and Rebecca M. Willett. Sparse Poisson intensity reconstruction algorithms. In *IEEE Workshop on Statistical Signal Processing*, 2009.

Zachary T. Harmany, Roummel F. Marcia, and Rebecca M. Willett. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms–theory and practice. *IEEE Transactions on Image Processing*, 21(3): 1084–1096, 2012.

H. Malcolm Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Annals of Statistics*, 6(3):473–484, 1978.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. arXiv: 2112.11079, 2021.

J. K. Lindsey. Comparison of probability distributions. *Journal of the Royal Statistical Society: Series B*, 36 (1):38–47, 1974.

J. K. Lindsey and G. Mersch. Fitting and comparing probability distributions with log linear models. *Computational Statistics & Data Analysis*, 13(4):373–384, 1992.

Florian Luisier, Cédric Vonesch, Thierry Blu, and Michael Unser. Fast interscale wavelet denoising of Poisson-corrupted images. *Signal Processing*, 90(2):415–427, 2010.

Colin Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. arXiv: 2301.07276, 2023.

Natalia L. Oliveira, Jing Lei, and Ryan J. Tibshirani. Unbiased risk estimation in the normal means problem via coupled bootstrap techniques. arXiv: 2111.09447, 2021.

Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259, 2006.

Maxim Raginsky, Rebecca M. Willett, Zachary T. Harmany, and Roummel F. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010.

Saharon Rosset and Ryan J. Tibshirani. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 15 (529):138–151, 2020.

Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local PCA. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.

Xiaotong Shen, Hsin-Cheng Huang, and Jimmy Ye. Adaptive model selection and assessment for exponential family distributions. *Technometrics*, 46(3):306–317, 2004.

Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6): 1135–1151, 1981.

Jianming Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

A Proofs and additional details for Section 2

A.1 Proof of Lemma 2

The joint probability mass function of (Y, ω) is given by

$$\begin{aligned}\mathbb{P}(Y = y, \omega = w) &= \mathbb{P}(\omega = k \,|\, Y = y)\mathbb{P}(Y = y) \\ &= \binom{y}{k} p^k (1-p)^{y-k} \frac{\mu^y e^{-\mu}}{y!},\end{aligned}$$

for all $y \geq 0$ and $k \in \{0, \ldots, y\}$. The probability mass function of $(U, V) = (Y - \omega, \omega)$ is thus

$$\begin{aligned}\mathbb{P}(U = u, V = v) &= \mathbb{P}(Y = u + v, \omega = v) \\ &= \binom{u+v}{v} p^v (1-p)^u \frac{\mu^{u+v} e^{-\mu}}{(u+v)!} \\ &= \frac{1}{u!v!} p^v (1-p)^u \mu^u \mu^v e^{-(p+1-p)\mu} \\ &= \frac{((1-p)\mu)^u e^{-(1-p)\mu}}{u!} \frac{(p\mu)^v e^{-p\mu}}{v!}.\end{aligned}$$

This shows that U and V are independent $\mathrm{Pois}((1-p)\mu)$ and $\mathrm{Pois}(p\mu)$ random variables, respectively, which proves the desired result.

A.2 Divergence deviance terms in UE versus CB

For deviance loss, a summand in the unbiased estimator (12) is undefined if $Y_i \neq 0$ and $g_i(Y - e_i) = 0$, and a summand in the CB estimator (19) is undefined if $Y_i^\dagger \neq 0$ and $g_i(Y^*) = 0$, where $Y^*, Y^\dagger$ are a sample from (16) (we have hidden the dependence on b). Suppose for simplicity that $n = 1$ and $g(0) = 0$. We can then compute, under $Y \sim \mathrm{Pois}(\mu)$, the probability with which this happens for the unbiased and CB estimators. For the unbiased estimator, this is:

$$\mathbb{P}(Y = 1) = e^{-\mu}\mu.$$

For the CB estimator, this is:

$$\begin{aligned}\sum_{y=1}^{\infty} \mathbb{P}(Y^* = 0 \,|\, Y = y)\mathbb{P}(Y = y) &= \sum_{y=1}^{\infty} \mathbb{P}(\omega = Y \,|\, Y = y)\mathbb{P}(Y = y) \\ &= \sum_{y=1}^{\infty} p^y e^{-\mu} \mu^y / y! \\ &= e^{-(1-p)\mu} \sum_{y=1}^{\infty} e^{-p\mu} (p\mu)^y / y! \\ &= e^{-(1-p)\mu}(1 - e^{-p\mu}) \\ &= e^{-\mu}(e^{p\mu} - 1).\end{aligned}$$

Figure 9 plots these two probabilities, a functions of μ, for p ranging over 0.01, 0.1, 0.3, 0.5. For small p, it is clear that the CB estimator has much lower probability of being ill-defined than the unbiased estimator.

A.3 Proof of Proposition 3

In this proof we use $f(Y_p, \tilde{Y}_p)$ to denote $D_\phi(\tilde{Y}_p, g(Y_p))$. First, we show that the map is continuous. For any $p \in [0, 1)$,

$$\lim_{t \to p} \mathbb{E}[f(Y_t, \tilde{Y}_t)] = \lim_{t \to p} \sum_{y_1, \ldots, y_n = 0}^{\infty} \sum_{\tilde{y}_1, \ldots, \tilde{y}_n = 0}^{\infty} f(y, \tilde{y}) \frac{e^{-2(1-t)\sum_{i=1}^n \mu_i} (\prod_{i=1}^n \mu_i^{y_i + \tilde{y}_i})(1-t)^{\sum_{i=1}^n y_i + \tilde{y}_i}}{\prod_{i=1}^n y_i! \tilde{y}_i!}$$

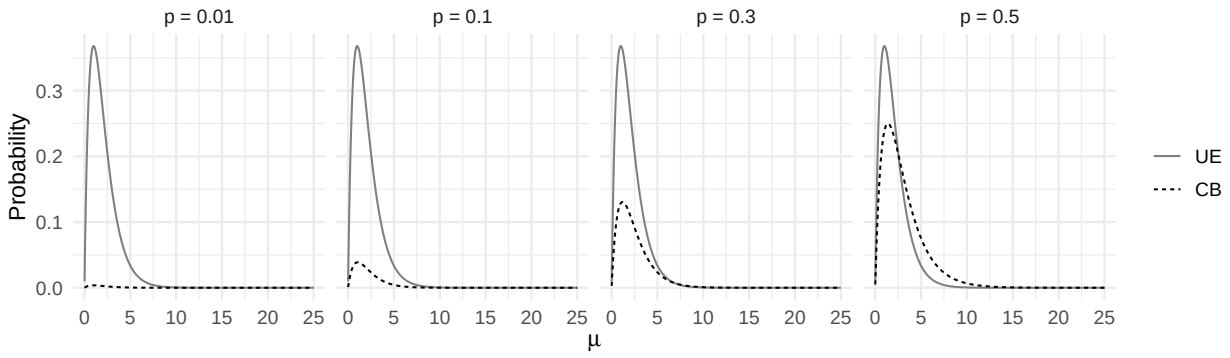


Figure 9: Comparison of the probabilities of an individual summand from the unbiased and CB estimators being ill-defined, as functions of the mean μ. The four panels show different values of p.

$$= \sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} f(y,\tilde{y}) \frac{e^{-2(1-p)\sum_{i=1}^n \mu_i} (\prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i})(1-p)^{\sum_{i=1}^n y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!}$$
$$= \mathbb{E}[f(Y_p, \tilde{Y}_p)].$$

To switch the infinite sum and the limit, we used the dominated convergence theorem (DCT). The dominating function is given by

$$h(\tilde{y}, y) = f(y,\tilde{y}) \frac{\prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!},$$

which is integrable by assumption:

$$\sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} h(\tilde{y}, y) = \sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} f(y,\tilde{y}) \frac{\prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!} e^{-2\sum_{i=1}^n \mu_i} e^{2\sum_{i=1}^n \mu_i}$$
$$= e^{2\sum_{i=1}^n \mu_i} \mathbb{E}[f(Y_0, \tilde{Y}_0)] < \infty.$$

Next, for the first derivative, note that

$$\frac{\partial}{\partial p}\mathbb{E}[f(Y_p, \tilde{Y}_p)] = \frac{\partial}{\partial p} \sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} f(y,\tilde{y}) \frac{e^{-2(1-p)\sum_{i=1}^n \mu_i} \prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i} (1-p)^{\sum_{i=1}^n y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!}$$
$$= \sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} \frac{f(y,\tilde{y}) \prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!} \frac{\partial}{\partial p} e^{-2(1-p)\sum_{i=1}^n \mu_i} (1-p)^{\sum_{i=1}^n y_i+\tilde{y}_i}$$
$$= \sum_{y_1,\ldots,y_n=0}^{\infty} \sum_{\tilde{y}_1,\ldots,\tilde{y}_n=0}^{\infty} \frac{f(y,\tilde{y}) \prod_{i=1}^n \mu_i^{y_i+\tilde{y}_i}}{\prod_{i=1}^n y_i!\tilde{y}_i!} \Bigg(2\sum_{i=1}^n \mu_i e^{-2(1-p)\sum_{i=1}^n \mu_i} (1-p)^{\sum_{i=1}^n y_i+\tilde{y}_i}$$
$$- \sum_{i=1}^n (y_i+\tilde{y}_i) e^{-2(1-p)\sum_{i=1}^n \mu_i} (1-p)^{\sum_{i=1}^n y_i+\tilde{y}_i-1} \Bigg)$$
$$= 2\sum_{i=1}^n \mu_i \mathbb{E}[f(Y_p, \tilde{Y}_p)] - \frac{1}{1-p}\mathbb{E}\big[f(Y_p, \tilde{Y}_p)\langle \tilde{Y}_p + Y_p, 1_n\rangle\big],$$

where we used DCT to switch the sums and derivative, using a similar dominating function as above and recognizing that the summand is Lipschitz in p with Lipschitz constant depending on $\mu, y, \tilde{y}$. Now to prove continuity of the first derivative, we apply the above continuity result with $f(Y_p, \tilde{Y}_p)\langle \tilde{Y}_p + Y_p, 1_n\rangle$ in place of $f(Y_p, \tilde{Y}_p)$. For k^{th} derivatives, the argument follows from sequential applications of the same continuity result and similar derivative calculations.

B Proof of Theorem 1

We start with a lemma that contains two key results to be used in the proof of Theorem 1.

Lemma 3. *Let $h : \mathbb{Z}_+^n \to \mathbb{R}_+^n$, and set $h(z) = 0$ for $z \notin \mathbb{Z}_+^n$. Fix any $y \in \mathbb{Z}_+^n$, and draw $\omega_i \sim \text{Binom}(y_i, p)$, independently, for $i = 1, \ldots, n$, where $p \in [0, 1)$. Then, for each $i = 1, \ldots, n$,*

(a) $\lim_{p\to 0} \mathbb{E}[h_i(y - \omega)] = h_i(y)$;

(b) $\lim_{p\to 0} \frac{1-p}{p} \cdot \mathbb{E}[\omega_i h_i(y - \omega)] = y_i h_i(y - e_i)$.

Recall, we use $e_i \in \mathbb{R}^n$ to denote the vector whose i^{th} entry is 1, with all others 0.

Proof. Define the following sets:

$$\begin{aligned}
\Omega &= \{(\omega_1, \ldots, \omega_n) : \omega_j \in \{0, \ldots, y_j\}, j = 1, \ldots, n\},\\
\Omega_{\setminus 0} &= \Omega \setminus \{0\},\\
\Omega_{i0} &= \{(\omega_1, \ldots, \omega_n) : \omega_i = 0, \omega_j \in \{0, \ldots, y_j\}, j \neq i\},\\
\Omega_{i10} &= \{(\omega_1, \ldots, \omega_n) : \omega_i = 1, \omega_j = 0, j \neq i\},\\
\Omega_{i1\bar{0}} &= \{(\omega_1, \ldots, \omega_n) : \omega_i \geq 1, \omega_j \in \{0, \ldots, y_j\}, j \neq i\} \setminus \Omega_{i10}.
\end{aligned}$$

For the first result (a), we have

$$\begin{aligned}
\lim_{p\to 0} \mathbb{E}[h_i(y - \omega)] &= \lim_{p\to 0} \sum_{\omega\in\Omega} h_i(y - \omega) p^{\sum_{k=1}^n \omega_k} (1 - p)^{\sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&= \lim_{p\to 0} h_i(y)(1 - p)^{\sum_{k=1}^n y_k}\\
&\quad + \sum_{\omega\in\Omega_{\setminus 0}} \lim_{p\to 0} h_i(y - \omega) p^{\sum_{k=1}^n \omega_k} (1 - p)^{\sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&= h_i(y),
\end{aligned}$$

since for $\omega \in \Omega_{\setminus 0}$, we have that $\sum_{k=1}^n \omega_k > 0$. For the second result (b),

$$\begin{aligned}
\lim_{p\to 0} \frac{1-p}{p} \mathbb{E}[\omega_i h_i(y - \omega)] &= \lim_{p\to 0} \frac{1-p}{p} \sum_{\omega\in\Omega} \omega_i h_i(y - \omega) p^{\sum_{k=1}^n \omega_k} (1 - p)^{\sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&= \lim_{p\to 0} \sum_{\omega\in\Omega} \omega_i h_i(y - \omega) p^{\sum_{k=1}^n \omega_k - 1} (1 - p)^{1 + \sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&= \lim_{p\to 0} \sum_{\omega\in\Omega_{i10}} \omega_i h_i(y - \omega) p^{\sum_{k=1}^n \omega_k - 1} (1 - p)^{1 + \sum_{k=1}^n Y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&= \lim_{p\to 0} h_i(y - e_i) p^0 (1 - p)^{1 + \sum_{k=1}^n y_k - \omega_k} \binom{y_i}{1} \prod_{j\neq i, j=1}^n \binom{y_j}{0}\\
&= y_i h_i(y - e_i),
\end{aligned}$$

where we use the fact that $\Omega = \Omega_{i0} \cup \Omega_{i10} \cup \Omega_{i1\bar{0}}$ and

$$\begin{aligned}
&\lim_{p\to 0} \sum_{\omega\in\Omega_{i0}} \omega_i h_i(y - \omega) p^{\sum_{k=1}^n \omega_k - 1} (1 - p)^{1 + \sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}\\
&\qquad = \lim_{p\to 0} \sum_{\omega\in\Omega_{i0}} 0 h_i(y - \omega) p^{\sum_{k=1}^n \omega_k - 1} (1 - p)^{1 + \sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j} = 0
\end{aligned}$$

as well as

$$\lim_{p\to 0} \sum_{\omega\in\Omega_{i1\bar{0}}} \omega_i h_i(y-\omega) p^{\sum_{k=1}^n \omega_k - 1}(1-p)^{1+\sum_{k=1}^n y_k - \omega_k} \prod_{j=1}^n \binom{y_j}{\omega_j}$$
$$= \sum_{\omega\in\Omega_{i1\bar{0}}} \omega_i h_i(y-\omega) 0^{\sum_{k=1}^n \omega_k - 1} \prod_{j=1}^n \binom{y_j}{\omega_j} = 0,$$

since for $\omega \in \Omega_{i1\bar{0}}$, we have that $\sum_{k=1}^n \omega_k - 1 > 0$. □

Now we are ready to prove Theorem 1. We start by expanding the infinite bootstrap estimator

$$\begin{aligned}
\mathbb{E}[\mathrm{CB}_p(g) \,|\, Y] &= \sum_{i=1}^n \mathbb{E}\Big[D_\phi(Y_i^\dagger, g_i(Y^*)) + \phi(Y_i^*) - \phi(Y_i^\dagger) \,\Big|\, Y\Big] \\
&= \sum_{i=1}^n \mathbb{E}\Big[\phi(Y_i^*) - \phi(g_i(Y^*)) - \nabla_i\phi(g(Y^*))(Y_i^\dagger - Y_i^*) \,\Big|\, Y\Big] \\
&= \sum_{i=1}^n \mathbb{E}\left[\phi(Y_i - \omega_i) - \phi(g_i(Y-\omega)) - \nabla_i\phi(g(Y-\omega))\left(\frac{1-p}{p}\omega_i - (Y_i - \omega_i)\right) \,\middle|\, Y\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\phi(Y_i - \omega_i) - \phi(g_i(Y-\omega)) - \frac{1-p}{p}\nabla_i\phi(g(Y-\omega))\omega_i + \nabla_i\phi(g(Y-\omega))(Y_i - \omega_i) \,\middle|\, Y\right].
\end{aligned}$$

Then, taking the limit in the last line,

$$\begin{aligned}
&\sum_{i=1}^n \lim_{p\to 0} \mathbb{E}\left[\phi(Y_i - \omega_i) - \phi(g_i(Y-\omega)) - \frac{1-p}{p}\nabla_i\phi(g(Y-\omega))\omega_i + \nabla_i\phi(g(Y-\omega))(Y_i - \omega_i) \,\middle|\, Y\right] \\
&\quad= \sum_{i=1}^n \lim_{p\to 0} \mathbb{E}\Big[\phi(Y_i - \omega_i) - \phi(g_i(Y-\omega)) + \nabla_i\phi(g(Y-\omega))(Y_i - \omega_i) \,\Big|\, Y\Big] - Y_i\nabla_i\phi(g(Y-e_i)) \\
&\quad= \sum_{i=1}^n \phi(Y_i) - \phi(g_i(Y)) + \nabla_i\phi(g(Y))(Y_i) - Y_i\nabla_i\phi(g(Y-e_i)) \\
&\quad= \mathrm{UE}(Y),
\end{aligned}$$

where in the second-to-last and last lines we used Lemma 3 parts (b) and (a), respectively. This completes the proof.

C Proofs for Section 4

C.1 Proof of Proposition 4

From Proposition 3, the mapping $p \mapsto \mathrm{Err}_p(g)$ has a continuous derivative for $p \in [0,1)$. By an application of the fundamental theorem of calculus, we can write the bias as

$$\begin{aligned}
\mathrm{Err}_p(g) - \mathrm{Err}(g) &= \int_0^p \frac{\partial}{\partial t}\mathrm{Err}_t(g)\,dt \\
&= \int_0^p \left\{2\sum_{i=1}^n \mu_i \mathbb{E}[D_\phi(\tilde{Y}_t, g(Y_t))] - \frac{1}{(1-t)}\mathbb{E}\big[D_\phi(\tilde{Y}_t, g(Y_t))\langle \tilde{Y}_t + Y_t, 1_n\rangle\big]\right\} dt \\
&= \int_0^p \left\{2\sum_{i=1}^n \mu_i \mathrm{Err}_t(g) - \frac{1}{(1-t)}\mathrm{Cov}\Big(D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big) - \frac{1}{(1-t)}2\mathrm{Err}_t(g)(1-t)\sum_{i=1}^n \mu_i\right\} dt \\
&= -\int_0^p \frac{1}{(1-t)}\mathrm{Cov}\Big(D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big)\,dt
\end{aligned}$$

$$
\begin{aligned}
&= -\int_0^p \frac{1}{(1-t)} \mathrm{Cor}\Big(D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big) \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}_t, g(Y_t))\big]} \sqrt{\mathrm{Var}\big[\langle \tilde{Y}_t + Y_t, 1_n\rangle\big]}\, dt \\
&= -\int_0^p \frac{1}{(1-t)} \mathrm{Cor}\Big(D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big) \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}_t, g(Y_t))\big]} \sqrt{(2(1-t)\sum_{i=1}^n \mu_i}\, dt \\
&= -\int_0^p \frac{1}{\sqrt{1-t}} \mathrm{Cor}\big[D_\phi(\tilde{Y}_t, g(Y_t)), \langle \tilde{Y}_t + Y_t, 1_n\rangle\Big) \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}_t, g(Y_t))\big]}\, dt \cdot \sqrt{2\sum_{i=1}^n \mu_i}.
\end{aligned}
$$

which proves (24). Upper bounding the correlation by 1, and using the monotone variance assumption for $p \in [0, 1/2]$, we have

$$
\begin{aligned}
|\mathrm{Err}_p(g) - \mathrm{Err}(g)| &\le \int_0^p \frac{1}{\sqrt{1-t}} \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}_t, g(Y_t))\big)}\, dt \cdot \sqrt{2\sum_{i=1}^n \mu_i} \\
&\le \int_0^p \frac{1}{\sqrt{1-t}} \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}, g(Y))\big)}\, dt \cdot \sqrt{2\sum_{i=1}^n \mu_i} \\
&= \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}, g(Y))\big)} \sqrt{2\sum_{i=1}^n \mu_i} \int_0^p \frac{1}{\sqrt{1-t}}\, dt \\
&= \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}, g(Y))\big)} \sqrt{2\sum_{i=1}^n \mu_i} \frac{2p}{1+\sqrt{1-p}} \\
&= \sqrt{\mathrm{Var}\big[D_\phi(\tilde{Y}, g(Y))\big)} \sqrt{2\sum_{i=1}^n \mu_i} \frac{5p}{3},
\end{aligned}
$$

which proves (25).

C.2 Proof of Proposition 5

We start from the fact that

$$
\mathrm{Var}[\mathrm{CB}_p \,|\, Y] = \frac{1}{B}\mathrm{Var}\Big[D_\phi(Y^\dagger, g(Y^*)) + \phi(Y^*) - \phi(Y^\dagger) \,\Big|\, Y\Big].
$$

Then

$$
\begin{aligned}
\mathbb{E}\big[\mathrm{Var}[\mathrm{CB}_p \,|\, Y]\big] &= \frac{1}{B}\mathbb{E}\Big[\mathrm{Var}\Big[D_\phi(Y^\dagger, g(Y^*)) + \phi(Y^*) - \phi(Y^\dagger) \,\Big|\, Y\Big]\Big] \\
&= \frac{1}{B}\mathbb{E}\Big[\mathrm{Var}\Big[\phi(Y^*) - \phi(g(Y^*)) - \langle \nabla\phi(g(Y^*)), Y^\dagger - g(Y^*)\rangle \,\Big|\, Y\Big]\Big] \\
&= \frac{1}{B}\mathbb{E}\Big[\mathrm{Var}\Big[\phi(Y^*) - \phi(g(Y^*)) - \langle \nabla\phi(g(Y^*)), Y^\dagger - Y^* + Y^* - g(Y^*)\rangle \,\Big|\, Y\Big]\Big] \\
&= \frac{1}{B}\mathbb{E}\Big[\mathrm{Var}\Big[D_\phi(Y^*, g(Y^*)) - \langle \nabla\phi(g(Y^*)), Y^\dagger - Y^*\rangle \,\Big|\, Y\Big]\Big] \\
&= \frac{1}{B}\mathbb{E}\Big[\mathrm{Var}\Big[D_\phi(Y^*, g(Y^*)) - \frac{1}{p}\langle \omega, \nabla\phi(g(Y^*))\rangle + \langle \nabla\phi(g(Y^*)), Y\rangle \,\Big|\, Y\Big]\Big] \\
&\le \frac{2}{B}\mathbb{E}\Big[\mathrm{Var}\Big[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y\rangle \,\Big|\, Y\Big]\Big] + \frac{2}{Bp^2}\mathbb{E}\Big[\mathrm{Var}\Big[\langle \omega, \nabla\phi(g(Y^*))\rangle \,\Big|\, Y\Big]\Big] \\
&\le \frac{2}{B}\mathrm{Var}\Big[D_\phi(Y^*, g(Y^*)) + \langle Y^*, \nabla\phi(g(Y^*))\rangle\Big] + \frac{2}{Bp^2}\mathrm{Var}\Big[\langle \omega, \nabla\phi(g(Y^*))\rangle\Big]
\end{aligned}
$$

$$= \frac{2}{B}\mathrm{Var}\Big[D_\phi(Y, g(Y)) + \langle Y, \nabla\phi(g(Y))\rangle\Big] + \frac{2}{B}O(p) + \frac{2}{Bp^2}\mathrm{Var}\big[\langle \omega, \nabla\phi(g(Y^*))\rangle\big],$$

where the second-to-last line uses the law of total variance, and the last line uses a Taylor expansion which follows from the continuity in p of expected value of functions of Y_p as assumed in the proof of Proposition 3. For the last term in the above display, note that ω and Y^* are independent. Therefore, we can use that fact that if X_1 and X_2 are independent, then $\mathrm{Var}[X_1X_2] = \mathrm{Var}[X_1]\mathrm{Var}[X_2] + \mathrm{Var}[X_1]\mathbb{E}^2[X_2] + \mathrm{Var}[X_2]\mathbb{E}^2[X_1]$, which translates to

$$\begin{aligned}
&\frac{2}{Bp^2}\mathrm{Var}\big[\langle \omega, \nabla\phi(g(Y^*))\rangle\big] \\
&\qquad = \frac{2}{Bp^2}\sum_{i=1}^n \mathrm{Var}\big[\omega_i \nabla_i\phi(g(Y^*))\big] \\
&\qquad = \frac{2}{Bp^2}\sum_{i=1}^n \Big(\mathrm{Var}[\omega_i]\mathrm{Var}\big[\nabla_i\phi(g(Y^*))\big] + \mathrm{Var}[\omega_i]\mathbb{E}^2\big[\nabla_i\phi(g(Y^*))\big] + \mathrm{Var}\big[\nabla_i\phi(g(Y^*))\big]\mathbb{E}^2[\omega_i]\Big) \\
&\qquad = \frac{2}{Bp^2}\sum_{i=1}^n \Big(p\mu_i\mathrm{Var}\big[\nabla_i\phi(g(Y^*))\big] + p\mu_i\mathbb{E}^2\big[\nabla_i\phi(g(Y^*))\big] + \mathrm{Var}\big[\nabla_i\phi(g(Y^*))\big]p^2\mu_i^2\Big) \\
&\qquad = \frac{2}{Bp}\sum_{i=1}^n \mu_i\mathbb{E}\big[\nabla_i\phi(g(Y^*))^2\big] + \frac{2}{B}\sum_{i=1}^n \mu_i^2\mathrm{Var}\big[\nabla_i\phi(g(Y^*))\big] \\
&\qquad = \frac{2}{Bp}\sum_{i=1}^n \mu_i\mathbb{E}\big[\nabla_i\phi(g(Y^*))^2\big].
\end{aligned}$$

Putting it all together gives the desired result (26).

C.3 Proof of Proposition 6

Note that the irreducible variance $\mathrm{Var}(\mathbb{E}[\mathrm{CB}_p(g)\,|\,Y])$ does not depend on B (because the inner expectation does not), so we assume without a loss of generality that $B = 1$ henceforth. Observe that

$$\begin{aligned}
&\mathrm{Var}\big(\mathbb{E}[\mathrm{CB}_p(g)\,|\,Y]\big) \\
&\qquad = \mathrm{Var}\Big(\mathbb{E}\Big[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y^* - Y^\dagger\rangle\,\Big|\,Y\Big]\Big) \\
&\qquad = \mathrm{Var}\left(\mathbb{E}\left[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y^*\rangle - \frac{1-p}{p}\langle \nabla\phi(g(Y^*)), \omega\rangle\,\middle|\,Y\right]\right) \\
&\qquad \le 2\mathrm{Var}\Big(\mathbb{E}\Big[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y^*\rangle\big|\,Y\Big]\Big) + 2\mathrm{Var}\left(\mathbb{E}\left[\frac{1-p}{p}\langle \nabla\phi(g(Y^*)), \omega\rangle\Big|\,Y\right]\right). \qquad (29)
\end{aligned}$$

For the first term in (29), we apply the law of total variance and

$$\begin{aligned}
\mathrm{Var}\Big(\mathbb{E}\Big[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y^*\rangle\Big]\Big) &\le \mathrm{Var}\Big[D_\phi(Y^*, g(Y^*)) + \langle \nabla\phi(g(Y^*)), Y^*\rangle\Big] \\
&\to \mathrm{Var}\Big[D_\phi(Y, g(Y)) + \langle \nabla\phi(g(Y)), Y\rangle\Big], \quad \text{as } p \to 0,
\end{aligned}$$

where the last convergence is guaranteed provided that $D_\phi^2(Y, g(Y))$ and $\langle \nabla\phi(g(Y)), Y\rangle^2$ have finite expectation, according to the first step in the proof of Proposition 3.

For the second term in (29) recalling that Φ_g as defined in the statement of the proposition,

$$\begin{aligned}
\lim_{p\to 0}\mathrm{Var}\left(\mathbb{E}\left[\frac{1-p}{p}\langle \nabla\phi(g(Y^*)), \omega\rangle\,\middle|\,Y\right]\right) &\le \lim_{p\to 0}\mathbb{E}\left(\mathbb{E}^2\left[\frac{1-p}{p}\langle \nabla\phi(g(Y^*)), \omega\rangle\,\middle|\,Y\right]\right) \\
&\le \lim_{p\to 0}\mathbb{E}\left(\mathbb{E}^2\left[\frac{1-p}{p}\langle \Phi_g(Y), \omega\,\middle|\,Y\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{p \to 0} (1-p)^2 \mathbb{E}\Big[\langle \Phi_g(Y), \mathbb{E}[\omega/p \,|\, Y] \rangle^2 \Big] \\
&= \lim_{p \to 0} (1-p)^2 \mathbb{E}\big[\langle \Phi_g(Y), Y \rangle^2 \big] \\
&= \mathbb{E}\big[\langle \Phi_g(Y), Y \rangle^2 \big].
\end{aligned}
$$

Putting it all together gives the desired result (27).