ON LIMITING THE PREDICTIVE EXPOSURE OF LEARNED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The outputs of machine learning algorithms can reveal properties of the underlying populations in the dataset. Some of these aggregate properties of the dataset may be considered sensitive or proprietary (such as the overall customer base and marketing strategy of a business, rather than a particular transaction). Thus, we want to limit the information leaked from a learning algorithm about a sensitive aggregate property of the unknown portion of the dataset. We formalize this leakage as predictive exposure. Given a portion of a dataset generated by a stochastic process, predictive exposure measures how well an observer can predict statistics about sensitive properties of the population of unknown individuals, before and after observing the algorithm's output. We show that when the leaked portion of the dataset is generated by a stationary, ergodic, mixing process, we can study the asymptotic predictive exposure of an algorithm using a classic tool, the generalized method of moments (Hansen, 1982). Our analysis shows how learning algorithms can be modified to limit predictive exposure. Using our analysis, we derive an exposure-limiting output perturbation scheme for ERM.

1 INTRODUCTION

Machine learning models are trained on large datasets containing contributions from many different individuals and organizations. The outputs of learned models can reveal the behaviors and tendencies of the underlying populations in the dataset. Some of these revealed trends may be considered proprietary or sensitive to various parties in the dataset. Consider a model trained on transactions between businesses and their customers. Each business has full knowledge of its own transactions, but has no knowledge of the transactions of other businesses. Aggregate properties of the unknown set of transactions, such as net revenue, average profit margin, and customer base, are considered highly proprietary, as these quantities reveal confidential business strategies and can affect competitive decisions. Thus, a business participating in the dataset may want assurance that the learned model does not substantially improve one business's ability to infer proprietary aggregate properties of other businesses. Concretely, consider when Business A contributes its transaction data to a shared dataset. The model provider uses this dataset to train a model to predict customer behavior. Based on its own transactions, Business A can estimate that $20\% \pm 5\%$ of the sales of other businesses in the dataset come from men under 30. However, after receiving the learned model, Business A can now sharpen its estimate to $27\% \pm 1\%$. Individual transactions may be protected, but the learned model has exposed a strategically sensitive population statistic of the remaining businesses. This example illustrates a form of information protection that is fundamentally different from protecting the information in individual records. Instead, we want to control how much a model release improves one's ability to predict specific sensitive statistical properties of the unknown population. More generally, given a compromised portion of the dataset, can we limit how much more information the output of an algorithm reveals about particular attributes of the unknown population?

In this work, we formalize this goal as *population predictive indistinguishability*: given a subset of the dataset, observing the output of an algorithm should not substantially improve the observer's ability to predict the protected statistic of the unknown population. We provide a framework for studying information leakage in settings where conventional notions of record-level protections are misaligned with the information that we actually want to protect. We quantify the population predictive indistinguishability of an algorithm through its *predictive exposure*. Given some initially revealed subset of the dataset, the predictive exposure measures how accurately an observer can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

predict a protected population statistic before and after observing the output of the algorithm. By studying predictive exposure, we design new mechanisms for protecting the sensitive properties of unknown populations. Our contributions are as follows.

- We introduce predictive exposure. Given a portion of a dataset generated by a stochastic process, the predictive exposure measures how well the observer can predict sensitive properties about the population of unknown individuals before and after observing the algorithm.
- We introduce a framework using the generalized method of moments (GMM) to compute the asymptotic predictive exposure of an algorithm with respect to a specified query when the leaked portion of the dataset is generated by a known stationary, ergodic, strongly mixing process. We prove that asymptotic predictive exposure is governed by a canonical correlation term between the query and the moment condition of the algorithm.
- We derive a predictive exposure-calibrated output perturbation scheme for ERM. The scheme adapts the noise covariance to the loss gradients and the process generating the leaked subset.
- We theoretically establish the benefits of calibrated noise over isotropic noise for linear regression, and empirically validate these benefits for deep learning models with an experiment on TOFU.
- We extend the framework to measure asymptotic predictive exposure for families of queries and when the exact process generating the leaked portion of the dataset is unknown.

2 Formalizing Predictive Exposure

We consider an observer who would like to infer some protected property about the unknown population of a dataset. In particular, we consider an observer that has knowledge of some subset of the dataset before observing the algorithm's output. We call this leaked subset of the dataset the core knowledge of the observer. They could gain this knowledge by being one of the contributors to the dataset (as in the business transaction example), by compromising some portion of the dataset, or through some public data release. We model the exposure of this subset as a stochastic process that reveals a portion of the dataset to the observer. Let $\mathcal{X}$ be an instance space, and let $S = (x_1, \ldots, x_N) \in \mathcal{X}^N$ be a dataset over N data points.

Definition 1 (Core Knowledge)**.** *Let $\mathcal{P}$ be a stochastic process that, given a dataset S, generates a sample $x_1, \ldots, x_n$, such that each $x_i \in S$. The observer's core knowledge is $C(S) = \{x_1, \ldots, x_n\}$.*

Consider the business transaction example. Let one of the businesses be the observer of interest. This business has complete knowledge of the purchase transactions of its customers only. Assume that customers are equally likely to purchase from any of the businesses. Then the core knowledge of this business is a random subset of the entire dataset of transactions. This can be modeled by the stochastic process $\mathcal{P}_{\mathrm{RS}}$ that produces each x_i by uniformly sampling from S without replacement. In general, $\mathcal{P}$ can be much more complex. Certain customers could be biased towards purchasing from a particular business. Certain customers may be more likely to publicly expose their purchases. Each of these disclosure scenarios can be captured by a stochastic process.

Let $A : \mathcal{X}^N \mapsto \Delta(\mathcal{F})$ be a randomized algorithm that takes in a dataset S and outputs a distribution over a model class $\mathcal{F}$. We assume that the designer of A knows the process $\mathcal{P}$ but does not know the realized core knowledge $C(S)$. This captures a setting where the algorithm designer knows how portions of the dataset could be leaked but does not know the exact set of compromised data. In Section 6, we relax the assumption that the process is known to the algorithm designer. The algorithm designer wants to protect the observer from accurately predicting some sensitive property. This is specified by a query q. Using their core knowledge $C(S)$ and the algorithm's output $A(S)$, the observer will try to predict the value of q on the unknown portion of the dataset as best they can.

Definition 2 (Bayes Optimal)**.** *Let $\Pi = \mathrm{Unif}(S \setminus C(S))$. Given knowledge $K(S)$ and query q, the Bayes optimal estimator for loss $\ell : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$ is $\hat{\theta}^*_{q|K(S)} = \arg\min_{\hat{\theta}} \mathbb{E}_{x\sim\Pi}[\ell(\hat{\theta}, q(x)) \mid K(S)]$.*

Since the observer wants to predict population properties about the unknown portion of the dataset, we measure the observer's prediction with respect to the average value of q on the unknown portion $S \setminus C(S)$. The Bayes optimal estimator represents the most accurate prediction of the average value of $q(x)$ that the observer could make. We primarily use $K(S) = C(S)$ or $K(S) = (C(S), A(S))$ to compare the difference in prediction before and after observing $A(S)$. The information contained in $C(S)$ about q is already revealed, so the best the algorithm can do is to ensure that the observer

does not gain any additional information about q through the algorithm's output. This additional information is measured through the predictive exposure.

Definition 3 (Predictive Exposure)**.** *Let process $\mathcal{P}$ produce the core knowledge of the observer $C(S)$. Let $q : \mathcal{X} \mapsto \mathcal{Y}$ be a sensitive query. Let $\ell : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$ be a loss function. Let $\hat{\theta}^*_{q|C(S)}$ and $\hat{\theta}^*_{q|C(S),A(S)}$ be the Bayes optimal estimator for $q(x)$ given $C(S)$, and given $C(S)$ and $A(S)$, respectively. Then algorithm A has $\gamma(\mathcal{P}, q, \ell, A)$-predictive exposure γ if*

$$\gamma(\mathcal{P}, q, \ell, A) = \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A} \Big[\mathop{\mathbb{E}}_{x\sim\Pi} [\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))] \Big] \leq \gamma.$$

Given core knowledge $C(S)$, the predictive exposure measures the increase in accuracy of the observer's prediction for the sensitive query q on the unknown population before and after observing the algorithm's output. If the difference in prediction is small, the algorithm's output reveals very little additional information beyond what an observer could have already learned from $C(S)$ about q. Predictive exposure as a measure of the additional information leakage by $A(S)$ about the sensitive query q on the unknown population is formalized when ℓ is the log loss.

Proposition 1. *If $q : \mathcal{X} \to [K]$ is a categorical query and $\ell(\hat{\theta}, y) = -y_i \log(\hat{\theta}_i)$ is the log loss, then predictive exposure $\gamma(\mathcal{P}, q, \ell, A) = \sup_S \mathrm{MI}(q(x); A(S) \mid C(S))$.*

For settings where we want to limit the predictive exposure for multiple sensitive properties, we extend our framework to bound $\gamma(\mathcal{P}, \mathcal{Q}, \ell, A) = \sup_{q\in\mathcal{Q}} \gamma(\mathcal{P}, q, \ell, A)$ for a family of queries $\mathcal{Q}$ in Section 6. In Appendix A.4, we extend the definition of predictive exposure beyond the estimation of $q(x)$ for the average unknown individual (in particular, the observer estimates $\mathbb{E}_{x\sim D'}[q(x)]$ for some distribution $D' \neq \Pi$ on the dataset). This extension allows us to reason about the predictive exposure of specific subpopulations of the unknown population. In Appendix A.8, we establish post-processing and non-adaptive composition properties of predictive exposure, analogous to the properties of the conditional mutual information framework from Steinke & Zakynthinou (2020).

3 Related Work

A broad line of work has studied how to limit what can be inferred from the outputs of algorithms about the underlying dataset, primarily through the lens of providing privacy for sensitive information. Prior works have largely considered two types of sensitive information: the information associated with individual records and global properties of the underlying dataset.

Protecting individual-level privacy. This has been extensively studied through differential privacy (DP) (Dwork et al., 2006; Dwork, 2008) and its variants (Cuff & Yu, 2016; Blum et al., 2008; Dong et al., 2019; Ghazi et al., 2021; Hall et al., 2011). DP ensures that, given arbitrary auxiliary information about the dataset, observing an algorithm's output does not reveal the presence of any individual record in the dataset. Differentially private algorithms achieve this goal by injecting noise. These algorithms can be implemented efficiently (Chaudhuri et al., 2011; Abadi et al., 2016).

Protecting global attributes of dataset. Beyond information associated with individual records, a related line of work considers settings where the sensitive information is a property of the dataset or a parameter of the data-generating distribution. Such frameworks include attribute privacy (Zhang et al., 2022), distribution privacy (Chen & Ohrimenko, 2023; Kawamoto & Murakami, 2019), inferential privacy (Ghosh & Kleinberg, 2017), and statistic maximal leakage (Wang et al., 2024). These frameworks specify a family of prior distributions over the dataset and measure how much information the algorithm's output leaks about a global dataset parameter under a specified family of priors. The differing goals of individual-level versus population-level privacy lead to different privacy mechanisms. However, these frameworks and their corresponding mechanisms often require strong assumptions on the family of priors to be implemented efficiently (Zhang et al., 2022; Song et al., 2017; Kawamoto & Murakami, 2019). Some privacy frameworks, such as Pufferfish privacy (Kifer & Machanavajjhala, 2014), Blowfish privacy (He et al., 2014), and g-leakage (Alvim et al., 2012), can be formulated to measure individual-level privacy or privacy of global attributes. These frameworks also specify a family of prior distributions over the dataset in order to measure information leakage and require significant assumptions on the family of priors to compute the leakage efficiently (Song et al., 2017; Nuradha & Goldfeld, 2023; Alvim et al., 2012; Wang et al., 2024).

The goals of population predictive indistinguishability and predictive exposure align with the latter line of work protecting global attributes. Although closely related to these prior works, predictive exposure differs in how it models the observer's initial knowledge. Rather than specifying a prior distribution over datasets, predictive exposure incorporates realized side information from the dataset directly through an exposed subset. This allows the framework to be applied to an arbitrary realized dataset without requiring the algorithm designer to posit a particular prior or distributional model, assumptions that may be difficult to specify or justify in practice. Furthermore, our proposed framework for calculating predictive exposure is efficient for any given dataset, making predictive exposure a practical empirical measure. Previous work on the privacy of global attributes has also established that the protection of global attributes and individual-level privacy notions studied by DP are incomparable (Zhang et al., 2022; Wang et al., 2024; Cormode, 2010). Each can exist without the other. We make the same observation for predictive exposure and differential privacy.

Proposition 2. *Let $\ell(\hat{\theta}, y) = (\hat{\theta} - y)^2$ be the squared loss. There exists a process $\mathcal{P}$, query q, and algorithm A such that A is ε-differentially private, $\varepsilon = O(\frac{1}{\sqrt{N}})$, but $\gamma(\mathcal{P}, q, \ell, A)$-predictive exposure is $\Omega(1)$. Conversely, there exists a process $\mathcal{P}$, query q, and algorithm A such that $\gamma(\mathcal{P}, q, \ell, A)$-predictive exposure is $O(\frac{1}{N})$, but A is not differentially private.*

Individual-level privacy can be strong, while predictive exposure is large. Predictive exposure can be small, while providing no individual-level privacy. We note that if we take the size of the revealed subset to $n = N - 1$, then predictive exposure against the set of all binary queries $\mathcal{Q}_{\text{bin}}$ under the log loss is equivalent to mutual information differential privacy (Cuff & Yu, 2016) (see Appendix A.7.3, Proposition 3). However, mutual information differential privacy provides relatively weak individual privacy guarantees, and predictive exposure is not designed to operate in this regime.

4 Computing Predictive Exposure

We focus on understanding the predictive exposure of large machine learning systems. Thus, we present a general framework to analyze the large-sample asymptotic predictive exposure of a query q when $n, N \to \infty$. To bound predictive exposure, we must bound the difference in loss between two estimators of q. For a convex and smooth loss (i.e. log loss or squared loss), the difference in the loss of the estimators can be characterized by the difference of the estimators themselves.

Lemma 1. *Let $\ell(\hat{\theta}) = \mathbb{E}_{x\sim\Pi}[\ell(\hat{\theta}, q(x)]$ be a convex, H-smooth loss. Let $\theta_0^q = \mathbb{E}_{x\sim\Pi}[q(x)]$. Then for any pair of estimators $\hat{\theta}_{q|C(S)}$ and $\hat{\theta}_{q|C(S),A(S)}$,*

$$\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}_{q|C(S)}, q(x)) - \ell(\hat{\theta}_{q|C(S),A(S)}, q(x))] \le H|\hat{\theta}_{q|C(S)} - \theta_0^q| \cdot |\hat{\theta}_{q|C(S)} - \hat{\theta}_{q|C(S),A(S)}|.$$

For log loss, if the true $p_0 = \Pr[q(x) = 1]$ satisfies $p_0 \in [\tau, 1-\tau]$, $H = \frac{1}{\tau^2}$. For squared loss, $H = 2$.

In particular, the difference in loss can be upper-bounded by the variance of $\hat{\theta}_{q|C(S)}$ and $\hat{\theta}_{q|C(S),A(S)}$ around the true θ_0^q. Thus, the best an observer can do is select the estimator with the smallest variance among all possible estimators. In statistics, this is known as the efficient estimator, representing the most accurate prediction that an observer can make, given infinite computational resources. We will find that, asymptotically, the efficient estimator coincides with the Bayes optimal estimator. Importantly, we do not assume that the observer uses the efficient estimator; it simply allows us to measure the worst-case increase in predictive exposure. If the variance of the efficient estimator given $C(S)$ and given $C(S), A(S)$ is known, then $|\hat{\theta}_{q|C(S)} - \theta_0^q|$ and $|\hat{\theta}_{q|C(S)} - \hat{\theta}_{q|C(S),A(S)}|$ can be bounded. Receiving $C(S)$ is equivalent to receiving a sample from a stochastic process on dataset S. If the process is stationary, ergodic, and mixing, and the information from $A(S)$ can be captured by a moment condition, the variance of the efficient estimators can be found through GMM.

4.1 Generalized Method of Moments

The generalized method of moments (GMM) (Hansen, 1982; 2010; Zsohar, 2012) is a semiparametric method for parameter estimation. Given a sample and known population moment conditions, GMM derives the efficient estimator for the parameter. GMM makes no assumptions on the data distribution. GMM estimators are consistent, asymptotically normal, and most efficient in the class

of estimators that do not use any information other than the moment conditions. We observe that the algorithm's output reveals a moment condition on the dataset. Consider when $A(S)$ is the mean of S. The mean is a population moment condition on the dataset. For a binary q under the log loss, the observer wants to estimate $p = \Pr[q(x) = 1]$. Given only the compromised subset, the observer can estimate p to some accuracy. The additional mean moment condition can improve the accuracy of the estimate. The improvement in the variance of the best estimate of p with and without the moment condition can be determined using GMM. GMM assumes that the sample is produced by a stationary, ergodic, α-mixing process. Some examples include i.i.d. sampling, sampling without replacement, weighted sampling, or sampling from an aperiodic, irreducible Markov chain.

Assumption 1. *$\mathcal{P}$ is a stationary, ergodic, and α-mixing process over N elements (see definitions in Appendix A.2). $\mathcal{P}$ has stationary distribution $D = (p_1, \ldots, p_N)$, such that $p_i > 0$, $\forall i \in [N]$.*

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from a stochastic process satisfying Assumption 1. Let $\theta \in \mathbb{R}^r$ be an unknown vector of parameters to estimate, with true value θ_0. Let population moment conditions $f(x, \theta)$ be a continuous and differentiable $\mathbb{R}^r \mapsto \mathbb{R}^m$ function of θ such that

$$\mathbb{E}[f(x, \theta)] = 0, \text{ only at the true parameter } \theta_0.$$

The first moment condition specifies the parameter of estimation, and additional moment conditions specify population parameters. Additional moment conditions have the form $g(x, \lambda)$ for some (possibly vector-valued) function g and population parameter λ. For example, consider the problem of estimating parameter p with true value $p_0 = \mathbb{E}[q(x)]$ for some query q, given a known population mean μ_0. We have $\theta = (p)$ and $f(x, \theta) = [q(x) - p,\ x - \mu_0]^\intercal$ where $g(x, \mu) = x - \mu_0$.

GMM also allows for noisy moment conditions. If instead of knowing λ directly, we only know $\tilde{\lambda} = \lambda_0 + \Delta$, where Δ is independent mean-zero noise. Then we have another unknown parameter λ with true value λ_0, $\theta = (p, \lambda)$, and moment conditions

$$f(x, \theta) = [q(x) - p,\ g(x, \lambda),\ \tilde{\lambda} - \lambda]^\intercal.$$

Let $f_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} f(x_i, \theta)$ be the sample moments. Given the sample, the GMM estimator of θ is

$$\hat{\theta} = \arg \min_{\theta \in \Theta \subseteq \mathbb{R}^r} f_n(\theta)^\intercal W f_n(\theta)$$

where W is some positive definite $m \times m$ weight matrix. Under standard regularity conditions (listed in Appendix A.9.2), we can characterize the behavior of the efficient GMM estimator $\hat{\theta}^{\text{GMM}}$.

Theorem 1 (Efficient GMM Estimator Hansen (1982))**.** *Let $G = \mathbb{E}[\frac{\partial f(x,\theta)}{\partial \theta}]$ be the Jacobian and let $\Omega = \sum_{k=-\infty}^{\infty} \mathbb{E}[f(x_0, \theta) f(x_k, \theta)^\intercal]$ be the long-run covariance. As $n \to \infty$, with $W = \Omega^{-1}(\theta_0)$, the efficient GMM estimator $\hat{\theta}^{\text{GMM}}$ is regular and asymptotically efficient with asymptotic distribution*

$$\sqrt{n}(\hat{\theta}^{\text{GMM}} - \theta_0) \sim \mathcal{N}(0, [G(\theta_0)^\intercal \Omega^{-1}(\theta_0) G(\theta_0)]^{-1}).$$

The proof of efficiency is standard from Hansen (1982). Importantly, the sample moments must follow a uniform law of large numbers (Theorem 16) and a central limit theorem (Theorem 17). From here, we also assume the posterior induced by both $C(S)$ and $C(S), A(S)$ satisfies a Bernstein–von Mises (BvM) theorem for the query functional. Under this assumption, the Bayes optimal estimators $\hat{\theta}^*_{q|C(S)}$ and $\hat{\theta}^*_{q|C(S),A(S)}$ are known to be regular and asymptotically efficient. We assume that the (functional) BvM holds, and under this condition, we show that one can replace the Bayes optimal estimator in the definition of predictive exposure with the asymptotically efficient GMM estimator. Standard sufficient conditions for the BvM theorem have been extensively studied in statistics literature (van der Vaart, 1998; Bickel et al., 1993; Ghosal & van der Vaart, 2017).

Theorem 2. *Let $\ell(\hat{\theta}) = \mathbb{E}_{x \sim \Pi}[\ell(\hat{\theta}, q(x)]$ be a twice continuously differentiable loss function with $\nabla \ell(\theta_0) = 0$ and $\nabla^2 \ell(\theta_0)$ positive definite. Then*

$$\mathop{\mathbb{E}}_{x \sim \Pi}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))] \leq \mathop{\mathbb{E}}_{x \sim \Pi}[\ell(\hat{\theta}^{\text{GMM}}_{q|C(S)}, q(x)) - \ell(\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}, q(x))] + o(\tfrac{1}{n})$$

Under our assumption, both estimators, $\hat{\theta}^{\text{GMM}}$ and $\hat{\theta}^*$, are regular and asymptotically efficient. Thus, both estimators must share the same influence function and therefore, the same behavior up to lower order terms. As $n \to \infty$, the lower order difference vanishes, and we can replace the difference in the loss of $\hat{\theta}^*$ with the difference in the loss of $\hat{\theta}^{\text{GMM}}$ in the definition of predictive exposure. Then we can use the behavior of $\hat{\theta}^{\text{GMM}}$ to bound the predictive exposure.

4.2 Applying GMM to Predictive Exposure

The observer receives $C(S)$, a random sample of S from $\mathcal{P}$. Using GMM for predictive exposure, we want to compare the behavior of the estimator given only $C(S)$ versus given $C(S)$ and $A(S)$. For GMM estimation, we first specify our parameter of interest. The observer wants to estimate the average value of $q(x)$ on the population; thus, $p = \mathbb{E}_{D_{\mathrm{U}}}[q(x)]$, $D_{\mathrm{U}} = \mathrm{Unif}(S)$. This is almost the right form for a moment condition, but for GMM estimation, we specify the moment condition under the expectation of the stationary distribution of the process D, rather than D_{U}. However, notice that a moment condition on D_{U} implies a moment condition on D. In particular, at the unique θ_0,

$$\mathbb{E}_{D_{\mathrm{U}}}[f(x,\theta_0)] = \mathbb{E}_{D}[w(x)f(x,\theta_0)] = 0, \text{ where } w(x_i) = \frac{dD_{\mathrm{U}}}{dD}(x_i) = \frac{1}{p_i N}.$$

To incorporate the information from the algorithm, consider $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, and $\Delta \sim \mathcal{N}(0, \Sigma_\Delta)$. $A(S)$ exactly reveals a (noisy) moment condition λ_0 for the dataset S. GMM can incorporate noisy or exact versions of $A(S)$. If we can express λ_0 in the form $\mathbb{E}[g(x,\lambda_0)] = \frac{1}{N}\sum_{i=1}^{N} g(x_i,\lambda_0) = 0$ for some g, then the information from $A(S)$ can be expressed as a moment condition on D_{U} and thus, a weighted moment condition on D. Then the moment conditions for predictive exposure are

$$\begin{aligned} f_{C(S)}(x,\theta) &= [w(x)(q(x)-p)], && \theta = (p), \\ f_{C(S),A(S)}(x,\theta) &= [w(x)(q(x)-p),\ w(x)g(x,\lambda),\ \tilde{\lambda}-\lambda]^\top, && \theta = (p,\lambda), \end{aligned} \tag{1}$$

where λ is an unknown parameter with true value λ_0. We note that θ must capture all unknown parameters, but we care only about the variance of the estimate of p.

Recall from Lemma 1, to bound predictive exposure, we must characterize the behavior of $|\hat{\theta}^{\mathrm{GMM}}_{q|C(S)} - \theta^q_0|$ and $|\hat{\theta}^{\mathrm{GMM}}_{q|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}|$. Theorem 1 characterizes the behavior of $|\hat{\theta}^{\mathrm{GMM}}_{q|C(S)} - \theta^q_0|$. For the behavior of $|\hat{\theta}^{\mathrm{GMM}}_{q|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}|$, we employ one additional property of GMM estimators.

Theorem 3. *When the efficient GMM estimators $\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}$ and $\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}$ are constructed on the same sample, their difference has asymptotic distribution*

$$\hat{\theta}^{\mathrm{GMM}}_{q|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)} \sim \mathcal{N}(0, \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}] - \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}])$$

$\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}]$ and $\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}]$ are found by computing $\frac{1}{n}[G^\top \Omega^{-1} G]^{-1}$ with the moment conditions in Equation 1. Then we use the concentration of the estimators to bound predictive exposure.

Definition 4 (Canonical Correlation (Anderson, 2003))**.** *Define correlation $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$. The canonical correlation between a scalar random variable X and a vector random variable V is $\rho_{\mathrm{cc}}(X,V) = \max_{c\in\mathbb{R}^d} \rho(X, c^\top V)$, the max correlation between X and any linear projection of V.*

Theorem 4 (Asymptotic Predictive Exposure)**.** *Let $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. Assume process $\mathcal{P}$ satisfies Assumption 1 and the moment conditions in Equation 1 satisfy the regularity conditions (Appendix A.9.2). Let $\sigma_1 = \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}]$ and $\sigma_2 = \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}]$. As $n, N \to \infty$, $n/N \to 0$, for a convex, H-smooth loss, with probability $1-\delta$,*

$$\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}, q(x)) - \ell(\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}, q(x))]] \le \frac{H\log(4/\delta)}{n} \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}.$$

Let $q' = w(x)(q(x) - p_0)$ and let $g' = w(x)g(x,\lambda_0)$. Let

$$\begin{aligned} c_0(w,q,g,\Delta) = \mathrm{Var}[q']^{-1}\mathrm{Cov}(q',g')\mathrm{Var}[g']^{-1}\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]\left(\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right] + \Sigma_\Delta^{-1}\right)^{-1} \\ \cdot \mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathrm{Cov}(q',g'). \end{aligned}$$

Deriving the form of σ_1 and σ_2, we have that $A(S)$ has predictive exposure

$$\gamma(\mathcal{P},q,\ell,A) \le \frac{H(1+\log(4))}{n} \cdot \mathrm{Var}[q'] \cdot \sqrt{\rho_{\mathrm{cc}}(q',g')^2 - c_0(w,q,g,\Delta)}.$$

For typical choices of losses (log loss and squared loss), we can plug in $H = \frac{1}{\tau^2}$ or $H = 2$, appropriately. All expectations in Theorem 4 are weighted averages on the dataset according to D. Therefore, given a dataset, all terms (expectations, variances, covariances) in Theorem 4 can be computed efficiently. Importantly, p_0 and λ_0 are the true values of $q(x)$ and $A(S)$ on the dataset.

Remark 1. *The resulting expression appears unwieldy, but we break down the key terms.*

- $\rho_{\mathrm{cc}}(q', g')^2$ - *The canonical correlation captures how correlated q and $A(S)$ are. Intuitively, it quantifies the most information we can extract from $A(S)$ about $q(x)$. Given a dataset and process, ρ_{cc} can be efficiently computed,* $\rho^2_{\mathrm{cc}}(q', g') = \frac{1}{\mathrm{Var}[q']} \cdot \mathrm{Cov}(q', g')\mathrm{Var}[g']^{-1}\mathrm{Cov}(g', q')$.
- $c_0(w, q, g, \Delta)$ - *c_0 captures how adding noise affects predictive exposure. c_0 ranges from 0 to $\rho_{\mathrm{cc}}(q', g')^2$. Given a dataset and stationary distribution, c_0 can also be efficiently computed.*

These terms capture how the structure of the algorithm and noise affect predictive exposure. For example, let $\Sigma_\Delta = \sigma^2 I_d$, then consider:

Condition	Predictive Exposure	Comment
$\sigma^2 \to 0$	$\gamma = \frac{H(1+\log 4)}{n} \cdot \mathrm{Var}[q'] \cdot \lvert\rho_{\mathrm{cc}}(q', g')\rvert$	Noiseless case, $c_0 \to 0$
$\sigma^2 \to \infty$	$\gamma = 0$	Infinite noise, $c_0 \to \rho_{\mathrm{cc}}(q', g')^2$
$\rho_{\mathrm{cc}}(q', g') = 0$	$\gamma = 0$	No exposure; q and A are unrelated

Thus, we have a general recipe for computing asymptotic predictive exposure.

- For a (noisy) algorithm $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, express λ_0 in the form $\mathbb{E}[g(x, \lambda_0)] = 0$.
- Define moment conditions $f_{C(S)}(x, \theta)$ and $f_{C(S),A(S)}(x, \theta)$, accordingly.
- Compute the variance of the efficient GMM estimators and apply Theorem 4.

Random Subset is Compromised Consider the process $\mathcal{P}_{\mathrm{RS}}$ where a sample of size $n = \alpha N$ is drawn uniformly without replacement from the dataset. Here, the observer compromises a random subset of data. Under $\mathcal{P}_{\mathrm{RS}}$, the stationary distribution is $D = D_{\mathrm{U}}$, and $w(x) = 1$.

Corollary 1. *Let $A(S) = \tilde{\lambda} = \lambda + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. As $n, N \to \infty$, $n/N \to \alpha$, for a convex, H-smooth ℓ,* $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) \le \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(x), g(x, \lambda))^2 - c_0(1, q, g, \Delta)}$.

5 Analyzing the Predictive Exposure of Algorithms

For several important algorithms, we demonstrate how to express $A(S)$ as $\mathbb{E}[g(x, \lambda_0)] = 0$.

Dataset Statistics Consider when the algorithm releases a noisy statistic of the dataset, such as the proportion of smokers. Here, specifying $A(S)$ as a moment condition is straightforward. A statistic λ_0 has the form, $\lambda_0 = \frac{1}{N}\sum_{i=1}^N \Gamma(x_i)$. For example, if λ_0 is the mean, $\Gamma(x) = x$; if λ_0 is a count statistic, $\Gamma(x) = \mathbb{1}\{x \text{ is a smoker}\}$. Then $g(x, \lambda) = \Gamma(x) - \lambda$ ensures $\mathbb{E}[g(x, \lambda)] = 0$ only at $\lambda = \lambda_0$.

Theorem 5. *Let $A(S) = \tilde{\lambda} = \frac{1}{N}\sum_{i=1}^N \Gamma(x_i) + \Delta$, $\Delta \sim \mathcal{N}(0, \sigma_\Delta)$. Let $g(x, \lambda) = \Gamma(x) - \lambda$. For a convex, H-smooth ℓ,* $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) = \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(x)] \cdot \rho(q(x), \Gamma(x)) \cdot \sqrt{\frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\sigma_\Delta}}$.

When $g(x, \lambda) = \Gamma(x) - \lambda$, we obtain a simplified expression, $c_0(1, q, g, \Delta) = \rho(q(x), \Gamma(x))^2 \cdot \left(1 - \frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\sigma_\Delta}\right)$. We can clearly see how the noise variance affects c_0 and predictive exposure.

Empirical Risk Minimization Next, consider an algorithm that releases a noisy empirical risk minimizer (ERM). Many machine learning problems can be reduced to ERM; thus, analyzing the predictive exposure of ERM is crucial. Let dataset entries consist of $z = (x, y)$ pairs. Consider a loss function $\ell(z, w)$. The ERM satisfies $w_{\mathrm{ERM}} = \arg\min_w \frac{1}{N}\sum_{i=1}^N \ell(z_i, w)$. This condition cannot be directly written in the form $\mathbb{E}[f(z, \theta)] = 0$. However, at w_{ERM}, we have $\frac{1}{N}\sum_{i=1}^N \nabla\ell(z_i, w_{\mathrm{ERM}}) = \mathbb{E}[\nabla\ell(z, w_{\mathrm{ERM}})] = 0$. Thus, let $\lambda_0 = w_{\mathrm{ERM}}$ and $g(x, w_{\mathrm{ERM}}) = \nabla\ell(z, w_{\mathrm{ERM}})$. To ensure proper GMM estimation, we restrict our attention to strictly convex ℓ so that $\mathbb{E}[g(x, w_{\mathrm{ERM}})] = 0$ only at w_{ERM}.

Theorem 6. *Let $w_{\mathrm{ERM}} = \arg\min_w \frac{1}{N}\sum_{i=1}^N \ell(z_i, w)$, for ℓ strictly convex and $\mathbb{E}[\nabla^2\ell(z, w_{\mathrm{ERM}})]$ positive definite. Let $A(S) = w_{\mathrm{ERM}} + \Delta$, $\Delta \sim \mathcal{N}(0, \Sigma_\Delta)$. Let $g(z, w_{\mathrm{ERM}}) = \nabla\ell(z, w_{\mathrm{ERM}})$. For a convex, H-smooth ℓ,* $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) \le \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(z)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(z), g(z, w_{\mathrm{ERM}}))^2 - c_0(w, q, g, \Delta)}$.

To achieve small canonical correlation, $q(z)$ must be fairly unrelated to all coordinates of the loss gradient. *Interestingly, it is not enough for the loss itself and the query to be uncorrelated.* In Appendix A.3, we show how to express batch gradient descent (BGD) as a moment condition, allowing

us to handle the approximate ERM of an arbitrary loss computed by BGD. A full characterization of predictive exposure for the ERM of a nonconvex loss is an important future direction.

5.1 Exposure-Calibrated Noise Scheme for ERM

We could add isotropic noise to the ERM, $\Sigma_\Delta = \sigma^2 I_d$. Instead, we can adapt the noise to the specific ERM and the process generating the compromised data. Analyzing the resulting terms of the GMM analysis, we propose a new mechanism to limit predictive exposure of an ERM.

Theorem 7. *Let* $\nabla_z = \nabla \ell(z, w_{\text{ERM}})$ *and* $H_z = \nabla^2 \ell(z, w_{\text{ERM}})$. *Let* $A(S) = \tilde{w} = w_{\text{ERM}} + \Delta$,

$$\Delta \sim \mathcal{N}(0, \sigma^2 \mathbb{E}[w(x) H_z]^{-1} \mathbb{E}[w(x)^2 \nabla_z \nabla_z^\top] \mathbb{E}[w(x) H_z]^{-1}).$$

For a convex, H-smooth ℓ, $\gamma(\mathcal{P}, q, \ell, A) \leq \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \text{Var}[q'] \cdot \rho_{\text{cc}}(q', g') \cdot \sqrt{\frac{1}{\sigma^2+1}}$.

The calibrated noise adapts to the specific ERM and process, yielding a clear relationship between the desired level of predictive exposure and the required noise.

Linear Regression We instantiate the calibrated noise scheme for linear regression under the random subset process to understand the structure of the calibrated noise and to highlight the improved tradeoff between predictive exposure and accuracy. For linear regression under $\mathcal{P}_{\text{RS}}$, we have $\ell(z, w) = \frac{1}{2}(w^\top x - y)^2$ and $w(x) = 1$. Thus, we add noise $\Delta \sim \mathcal{N}(0, \sigma^2 \mathbb{E}[xx^\top]^{-1} \mathbb{E}[(w_{\text{ERM}}^\top x - y)^2 xx^\top] \mathbb{E}[xx^\top]^{-1})$. Analyzing the structure, we see that more noise is added in directions where w_{ERM} has a large loss or x's are sparse. When w_{ERM} has a small empirical loss, $\tilde{w}$ is less noisy for the same level of predictive exposure. We formalize this result below.

Theorem 8. *Let* $\tilde{w} = w_{\text{ERM}} + \Delta$, $\Delta \sim \mathcal{N}(0, \sigma^2 \cdot \mathbb{E}[xx^\top]^{-1} \mathbb{E}[(w_{\text{ERM}}^\top x - y)^2 xx^\top] \mathbb{E}[xx^\top]^{-1})$. *Then*

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] \leq \min\left\{\frac{\sigma^2(\max_i \|x_i\|^2)}{\lambda_{\min}(\mathbb{E}[xx^\top])} \cdot \hat{\mathcal{L}}_S(w_{\text{ERM}}), \sigma^2 d \cdot \hat{\mathcal{L}}_S^{\max}(w_{\text{ERM}})\right\}.$$

where $\hat{\mathcal{L}}_S(w_{\text{ERM}}) = \frac{1}{N}\sum_{i=1}^{N}(w_{\text{ERM}}^\top x_i - y_i)^2$ *and* $\hat{\mathcal{L}}_S^{\max}(w_{\text{ERM}}) = \max_i (w_{\text{ERM}}^\top x_i - y_i)^2$.

Lemma 2. *When* $\Delta \sim \mathcal{N}(0, \sigma^2 I_d)$, $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N}\|x_i\|^2\right)$.

When training loss, $\hat{\mathcal{L}}_S(w_{\text{ERM}})$ or $\hat{\mathcal{L}}_S^{\max}(w_{\text{ERM}})$, is small, calibrated noise changes the model's prediction less on average than isotropic noise. *The GMM framework informs us of an improved mechanism to limit predictive exposure.* In Appendix A.5, we apply the calibrated mechanism to a linear regression dataset and demonstrate the improvement in model utility over isotropic noise.

6 Extensions to a Family of Queries and Unknown Process

We can extend our predictive exposure framework to settings where the algorithm designer would like to protect multiple population properties or when the exact process $\mathcal{P}$ generating the observer's core knowledge is unknown. Let $\mathcal{Q}$ be a family of queries that we would like to protect, and let $\boldsymbol{\mathcal{P}}$ be a family of possible processes generating the core knowledge.

Finite Family of Queries and Processes Notice that for a particular process $\mathcal{P} \in \boldsymbol{\mathcal{P}}$, the predictive exposure depends on the weight function $w \in \mathcal{W}_{\boldsymbol{\mathcal{P}}}$ derived from the stationary distribution of $\mathcal{P}$. From Theorem 4, we know that for a fixed weight function w and query q, with high probability,

$$\mathbb{E}_{x\sim\Pi}[\ell(\hat{\theta}^{\text{GMM}}_{q|C(S)}, q(x)) - \ell(\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}, q(x))]] \text{ is small.}$$

From here, we can employ a union bound to show that for all $w \in \mathcal{W}_{\boldsymbol{\mathcal{P}}}$, $q \in \mathcal{Q}$, the above is small with high probability. Thus, the worst-case predictive exposure over all $\boldsymbol{\mathcal{P}}$ and $\mathcal{Q}$ is bounded.

Theorem 9. *Let* $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. *For a finite family of queries* $\mathcal{Q}$, *a finite family of processes* $\boldsymbol{\mathcal{P}}$ *satisfying Assumption 1, and a convex, H-smooth* ℓ, *the predictive exposure is* $\sup_{\mathcal{P}\in\boldsymbol{\mathcal{P}}} \gamma(\mathcal{P}, \mathcal{Q}, \ell, A) \leq \frac{H(1+\log(4|\mathcal{Q}\|\mathcal{W}_{\boldsymbol{\mathcal{P}}}|))}{n} \sup_{w\in\mathcal{W}_{\boldsymbol{\mathcal{P}}}} \sup_{q\in\mathcal{Q}} \text{Var}[q'] \sqrt{\rho_{\text{cc}}(q', g')^2 - c_0(w, q, g, \Delta)}$

Linear Query Family Consider the family $\mathcal{Q}_{\text{lin}} = \{q(x) = u^\top x \mid u \in \mathcal{U}_{\text{lin}} \subseteq \mathbb{R}^d\}$. For infinitely large families, we must rely on something stronger than the concentration of the estimators of single

queries. Consider $\hat{\theta}^{\text{GMM}}_{x|C(S)}$ and $\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)}$, estimators for $\mathbb{E}[x \mid C(S)]$ and $\mathbb{E}[x \mid C(S), A(S)]$. Using GMM, we know that $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu) \sim \mathcal{N}(0, \Sigma_1)$ and $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)} - \mu) \sim \mathcal{N}(0, \Sigma_2)$. If estimates of x concentrate well, then estimates of linear functions of x also concentrate well.

Theorem 10. *If $\mathcal{U}_{\text{lin}}$ is contained within the Euclidean ball of radius B, for a convex, H-smooth ℓ,* $\gamma(\mathcal{P}_{\text{RS}}, \mathcal{Q}_{\text{lin}}, \ell, A) \leq \frac{4HB(2+\log(2))}{(1-\alpha)\alpha N} \cdot \text{tr}(\text{Var}[x]) \cdot \sqrt{\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\text{tr}(\text{Var}[x])}}$.

Definition 5 (First Canonical Correlation from CCA)**.** *For two vector random variables, X and Y,* $\rho^1_{\text{cc}}(X,Y)^2 = \lambda_{\max}(\text{Var}[X]^{-1}\text{Cov}(X,Y)\text{Var}[Y]^{-1}\text{Cov}(Y,X))$.

We recover a dependence on a correlation quantity, now between x and $g(x,\lambda)$. Canonical correlation analysis (CCA) is a foundational technique in dimension reduction and multiview learning (Kakade & Foster, 2007; Chaudhuri et al., 2009; Sindhwani et al., 2005; Sridharan & Kakade, 2008), and we find its connection to predictive exposure compelling. Importantly, CCA, and thus $\rho^1_{\text{cc}}(x, g(x,\lambda))$, can be computed efficiently. For a general $\mathcal{U}_{\text{lin}}$, see the predictive exposure bound in Appendix A.11.2, which depends on the Gaussian complexity (Definition 14) of $\mathcal{U}_{\text{lin}}$.

Lipschitz Query Family We can apply a similar argument for Lipschitz query families.

Theorem 11. *Let query family $\mathcal{Q}_{\text{L}} = \{f(x) \mid f \text{ is } L\text{-Lipschitz}\}$. For a convex, H-smooth loss ℓ,* $\gamma(\mathcal{P}_{\text{RS}}, \mathcal{Q}_{\text{L}}, \ell, A) \leq \frac{4HL^2(2+\log(2))}{(1-\alpha)\alpha N} \cdot \text{tr}(\text{Var}[x]) \cdot \sqrt{\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\text{tr}(\text{Var}[x])}}$.

7 Experiments

Although the GMM assumptions do not hold for deep learning models, the exposure-calibrated noise for ERM still informs how to add noise to models to limit predictive exposure. We implement our exposure-calibrated mechanism on a Llama-3.2-1B-Instruct model fine-tuned on the TOFU dataset (Maini et al., 2024; Dorna et al., 2025; Meta AI, 2024) to empirically validate our theoretical results. The TOFU dataset is a set of QA pairs about fictional authors. We can think of the fine-tuned model as the ERM of some loss on the TOFU dataset. Thus, we can (approximately) compute the variance of the calibrated noise using loss gradients and apply the noise to the model weights.

Let the sensitive property be the gender of the author. We add calibrated noise and isotropic noise at various scales to the TOFU fine-tuned model and compare how well each mechanism obfuscates gender. To measure the information exposed about gender in the noisy model weights, we extract author embeddings from each noisy model. Then we train a small MLP on a random subset of embeddings to predict gender and test how well the MLP can predict gender on unseen embeddings. High accuracy on unseen authors implies that the noisy embeddings carries enough information about gender, and serves as a proxy for predictive exposure. We also measure the utility of each noisy model. We find that at low noise scales, calibrated noise and isotropic noise lead to comparable predictive exposure of gender. However, at high noise scales, for the same model utility, the calibrated noise has significantly less predictive exposure of gender compared to isotropic noise. Full experiment details are described in Appendix A.6.

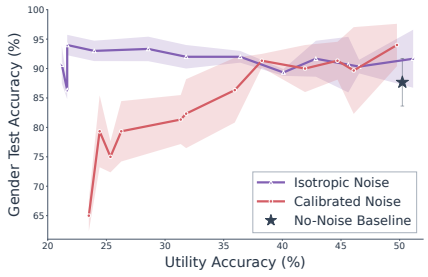


Figure 1: Plot of model utility versus the classifier's test accuracy on gender.

8 Conclusion

We present predictive exposure as a new measure for information leakage of a protected statistic of the unknown portion of the dataset through an algorithm's output. We present a general framework for computing the asymptotic predictive exposure using GMM. We find that the quantities underlying predictive exposure can be efficiently computed, giving algorithm designers a practical procedure to study information leakage of sensitive properties. Some open questions include characterizing the predictive exposure of the (approximate) ERM of a nonconvex loss, and developing adaptive composition for predictive exposure.

AI USE STATEMENT

In this work, we used generative AI tools to implement experiments (TOFU experiment and linear regression experiment). We have not used generative AI tools to develop theoretical models, prove mathematical claims, write proofs, propose hypotheses, or design research methodology. Generating synthetic datasets, cleaning datasets, data analysis, and translation assistance are not applicable to this work. Additionally, we used generative AI tools to identify relevant literature and edit writing. We have reviewed all AI-assisted work. The LLM-generated code and experiments were reviewed by the primary author. The primary author reviewed all AI-suggested references and writing edits. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI tools.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

M'rio S. Alvim, Kostas Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *2012 IEEE 25th Computer Security Foundations Symposium*, pp. 265–279, 2012. doi: 10.1109/CSF.2012.26.

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 3 edition, 2003.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, March 2003. ISSN 1532-4435.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pp. 609–618, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580470. doi: 10.1145/1374376.1374464. URL https://doi.org/10.1145/1374376.1374464.

Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 129–136, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553391. URL https://doi.org/10.1145/1553374.1553391.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011. ISSN 1532-4435.

Michelle Chen and Olga Ohrimenko. Protecting global properties of datasets with distribution privacy mechanisms. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7472–7491. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/chen23f.html.

Graham Cormode. Individual privacy vs population privacy: Learning to attack anonymization, 2010. URL https://arxiv.org/abs/1011.2511.

Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978308. URL http://dx.doi.org/10.1145/2976749.2978308.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy, 2019. URL https://arxiv.org/abs/1905.02383.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025. URL https://arxiv.org/abs/2506.12618.

Prosper Dovonon, Yves F. Atchadé, and Firmin Doko Tchatoka. Efficiency bounds for moment condition models with mixed identification strength. *Journal of Econometrics*, 248:105723, 2025. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2024.105723. URL https://www.sciencedirect.com/science/article/pii/S0304407624000691.

Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (eds.), *Theory and Applications of Models of Computation*, pp. 1–19, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-79228-4.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy, 2021. URL https://arxiv.org/abs/2102.06062.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms, 2017. URL https://arxiv.org/abs/1603.01508.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy, 2011. URL https://arxiv.org/abs/1112.2680.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

Lars Peter Hansen. *Generalized method of moments estimation*, pp. 105–118. Palgrave Macmillan UK, London, 2010. ISBN 978-0-230-28083-0. doi: 10.1057/9780230280830_13. URL https://doi.org/10.1057/9780230280830_13.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: tuning privacy-utility tradeoffs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD/PODS'14. ACM, June 2014. doi: 10.1145/2588555.2588581. URL http://dx.doi.org/10.1145/2588555.2588581.

Joseph Horowitz. A uniform law of large numbers and empirical central limit theorem for limits of finite populations. *Statistics & Probability Letters*, 10(2):159–166, 1990. ISSN 0167-7152. doi: https://doi.org/10.1016/0167-7152(90)90012-V. URL https://www.sciencedirect.com/science/article/pii/016771529090012V.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In Nader H. Bshouty and Claudio Gentile (eds.), *Learning Theory*, pp. 82–96, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72927-3.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/5b69b9cb83065d403869739ae7f0995e-Paper.pdf.

Yusuke Kawamoto and Takao Murakami. *Local Obfuscation Mechanisms for Hiding Probability Distributions*, pp. 128–148. Springer International Publishing, 2019. ISBN 9783030299590. doi: 10.1007/978-3-030-29959-0_7. URL http://dx.doi.org/10.1007/978-3-030-29959-0_7.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), January 2014. ISSN 0362-5915. doi: 10.1145/2514689. URL https://doi.org/10.1145/2514689.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024.

Daniel McFadden. *Econometrics 240B Second Half Reader*. 1999.

Meta AI. Llama 3.2: 1b instruct. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024.

Theshani Nuradha and Ziv Goldfeld. Pufferfish privacy: An information-theoretic study. *IEEE Transactions on Information Theory*, 69(11):7336–7356, November 2023. ISSN 1557-9654. doi: 10.1109/tit.2023.3296288. URL http://dx.doi.org/10.1109/TIT.2023.3296288.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics and Probability Letters, 33*, pp. 291–297, 1997.

C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975. doi: 10.1137/0712047. URL https://doi.org/10.1137/0712047.

Zuzana Prášková and Pranab Sen. Asymptotics in finite population sampling. *Handbook of Statistics*, 29:489–522, 12 2009. doi: 10.1016/S0169-7161(09)00240-5.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the Workshop on Learning with Multiple Views, 22nd ICML*, 2005.

Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data, 2017. URL https://arxiv.org/abs/1603.03977.

Karthik Sridharan and Sham Kakade. An information theoretic framework for multi-view learning. In *Conference on Learning Theory*, pp. 403–414, 01 2008.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information, 2020. URL https://arxiv.org/abs/2001.09122.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Shuaiqi Wang, Zinan Lin, and Giulia Fanti. Statistic maximal leakage, 2024. URL https://arxiv.org/abs/2411.18531.

Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 757–766, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533139. URL https://doi.org/10.1145/3531146.3533139.

Peter Zsohar. Short introduction to the generalized method of moments. *Hungarian Statistical Review, Special Number 16*, 2012.

A APPENDIX

CONTENTS

A.1 Notation 14
A.2 Definitions 14
A.3 Predictive Exposure of Batch Gradient Descent 15
A.4 Predictive Exposure Beyond Estimating the Average Value of $q(x)$ 16
A.5 Linear Regression Experiment 17
A.6 Details of TOFU Experiment 18
A.7 Proofs from Section 2 19
A.7.1 Proposition 1 19
A.7.2 Proposition 2 19
A.7.3 Proposition 3, Connection to Mutual Information Differential Privacy 21
A.8 Post-Processing and Composition of Predictive Exposure 23
A.9 Proofs from Section 4 23
A.9.1 Lemma 1 23
A.9.2 Regularity Conditions 24
A.9.3 Theorem 1, GMM Efficiency 25
A.9.4 Theorem 3, Difference of GMM Estimators 28
A.9.5 Theorem 4, Asymptotic Predictive Exposure of a Query 29
A.10 Proofs from Section 5 35
A.10.1 Theorem 5, Dataset Statistics 35
A.10.2 Theorem 6, Empirical Risk Minimization 36
A.10.3 Theorem 7, Calibrated Noise Scheme 37
A.10.4 Calibrated Noise and Linear Regression 37
A.11 Proofs from Section 6 39
A.11.1 Theorem 9, Finite Family of Queries and Processes 39
A.11.2 Theorem 10, Linear Queries 40
A.11.3 Theorem 11, Lipschitz Queries 46
A.12 Helpful Theorems 47

A.1 NOTATION

- $[N] = \{1, \dots, N\}$
- $\mathrm{Unif}(S)$ is the uniform distribution over the element in set S
- $X \sim \mathcal{N}(\mu, \Sigma)$ is normally distributed with mean μ and variance Σ
- $\mathbb{E}[X]$ is the expected value of X
- $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ for a one-dimensional random variable X
- $\mathrm{Var}[V] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]$ for a vector random variable V
- $\mathrm{Cov}(U, V) = \mathbb{E}[(U - \mathbb{E}[U])(V - \mathbb{E}[V])^\top]$
- $\rho(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X] \cdot \mathrm{Var}[Y]}}$ is the correlation between X and Y
- $D_{\mathrm{KL}}(P \| Q)$ is the KL-divergence between distributions P and Q
- $\mathrm{MI}(X; Y \mid Z)$ is the conditional mutual information between X and Y given Z
- $X_n \to_p Y_0$, X_n converges to Y_0 in probability
- $X_n \to_d Y_0$, X_n converges to Y_0 in distribution
- $A[i, j]$ denotes the (i, j) entry of matrix A

A.2 DEFINITIONS

Definition 6 ((ε, δ)-Differential Privacy). *A randomized algorithm A is (ε, δ)-differentially private, if for all neighboring datasets S, S', and for all $F \subseteq \mathcal{F}$,*

$$\Pr[A(S) \in F] \leq e^{\varepsilon} \Pr[A(S') \in F] + \delta.$$

Definition 7 (Stationarity). *The stochastic process $(X_t)_{t \in \mathbb{Z}}$ is stationary if for all $k \in \mathbb{N}$, all $t_1, \dots, t_k \in \mathbb{Z}$, and all $h \in \mathbb{Z}$, $(X_{t_1}, \dots, X_{t_k}) \overset{d}{=} (X_{t_1+h}, \dots, X_{t_k+h})$.*

Definition 8 (Ergodicity). *The stochastic process $(X_t)_{t \in \mathbb{Z}}$ is ergodic if for any integrable function f, the time average $\frac{1}{n} \sum_{t=1}^{n} f(X_t)$ converges almost surely to $\mathbb{E}[f(X_0)]$.*

Definition 9 (α-mixing). *For $k \geq 1$, the α-mixing coefficients are defined as*

$$\alpha(k) = \sup_{t \in \mathbb{Z}} \sup_{A \in \sigma(X_s : s \leq t), B \in \sigma(X_s : s \geq t+k)} |\Pr[A \cap B] - \Pr[A] \Pr[B]|.$$

The stochastic process $(X_t)_{t \in \mathbb{Z}}$ is α-mixing if $\alpha(k) \to 0$, as $k \to \infty$.

Definition 10 (Convex Function). *A function f is convex if $f(y) \geq f(x) + \nabla f(x)^\top \|y - x\|$.*

Definition 11 (Smooth Function). *A function f is H-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq H \cdot \|x - y\|$.*

Definition 12 (Lipschitz Function). *A function f is L-Lipschitz if $|f(x) - f(y)| \leq L \cdot \|x - y\|$.*

A.3 Predictive Exposure of Batch Gradient Descent

When the exact, unique ERM is released, we apply Theorem 6 to compute the predictive exposure. However, in many machine learning settings, the algorithm releases an approximate ERM, typically found through a version of gradient descent. We can express the result of batch gradient descent (BGD) as a moment condition. Recall the BGD update,

$$w_{t+1} = w_t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla \ell(z_i, w_t).$$

We observe that each step of BGD can be expressed as a moment condition,

$$g(z, w_{t-1,t}) = w_{t-1} - w_t - \eta \nabla \ell(z, w_{t-1}).$$

Given a sequence of w_t's from BGD on a dataset, we define the algorithm's moment condition $g(z, w_{1:T})$ as a vector containing all T steps of BGD.

$$g(z, w_{1:T}) = \begin{bmatrix} w_0 - w_1 - \eta \nabla \ell(z, w_0) \\ w_1 - w_2 - \eta \nabla \ell(z, w_1) \\ \vdots \\ w_{T-1} - w_T - \eta \nabla \ell(z, w_{T-1}) \end{bmatrix}$$

Thus, given the loss gradients at w_T and the sequence of w_t's, using Theorem 4, we can compute the predictive exposure for any approximate ERM obtained from BGD. Note that for a fixed starting point w_0, the sequence of w_t's that satisfy $\mathbb{E}[g(z, w_{1:T})] = 0$ is unique for any ℓ, thus eliminating the need to assume strictly convex ℓ for proper estimation.

A.4 Predictive Exposure Beyond Estimating the Average Value of $q(x)$

Consider the following extension of predictive exposure. We say algorithm A has $\gamma_{D'}(\mathcal{P}, \mathcal{Q}, \ell, A)$-predictive exposure γ if

$$\gamma_{D'}(\mathcal{P}, \mathcal{Q}, \ell, A) = \sup_S \sup_{q \in \mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}, A} [\mathop{\mathbb{E}}_{x\sim D'} [\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]] \le \gamma.$$

for some known distribution $D' = (p'_1, \ldots, p'_N)$ on dataset S. A simple example is when the attacker cares about the average of $q(x)$ on some particular subgroup $G \subseteq S$, and we have $D' = \mathrm{Unif}(G)$.

In Section 4.1, we demonstrate how to use the generalized method of moments to compute asymptotic $\gamma(\mathcal{P}, \mathcal{Q}, \ell, A)$-predictive exposure, where x is drawn uniformly from the set of unknown individuals, $x \sim \mathrm{II}$. We can compute asymptotic $\gamma_{D'}(\mathcal{P}, \mathcal{Q}, \ell, A)$-predictive exposure with one small change to the moment conditions.

Consider when the attacker wants to estimate $p = \mathbb{E}_{D'}[q(x)]$, $D' = (p'_1, \ldots, p'_N)$, given the information they have. Then, at the unique θ_0,

$$\mathbb{E}_{D'}[q(x) - p] = \mathbb{E}_D[w'(x)(q(x) - p)] = 0, \text{ where } w'(x_i) = \frac{dD_{\mathrm{U}}}{dD}(x_i) = \frac{p'_i}{p_i}.$$

We still expect $A(S)$ to have the form $\frac{1}{N}\sum_{i=1}^N g(x_i, \lambda) = 0$, so we maintain the same moment condition for $A(S)$,

$$\mathbb{E}_{\mathrm{D_U}}[g(x, \lambda)] = \mathbb{E}_D[w(x)g(x, \lambda)] = 0, \text{ where } w(x_i) = \frac{dD'}{dD}(x_i) = \frac{1}{p_i N}$$

Then we use the moment conditions

$$f_{C(S)}(x, \theta) = [w'(x)(q(x) - p)]$$
$$f_{C(S),A(S)}(x, \theta) = [w'(x)(q(x) - p),\ w(x)g(x, \lambda),\ \tilde{\lambda} - \lambda]^\top$$

to compute $\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}]$, $\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}]$, and the resulting predictive exposure.

Theorem 12. *Let $A(S) = \tilde{\lambda} = \lambda + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. Assume process $\mathcal{P}$ satisfies Assumption 1. Assume the moment conditions in Equation 1 satisfy the regularity conditions (Appendix A.9.2).*

Let $q'' = w'(x)(q(x) - p)$ and let $g' = w(x)g(x, \lambda)$. Then A has predictive exposure

$$\gamma(\mathcal{P}, q, \ell, A) \le \frac{H(1 + \log(4))}{n} \cdot \mathrm{Var}[q''] \cdot \sqrt{\rho_{\mathrm{cc}}(q'', g')^2 - c_0(w, q, g, \Delta)},$$

where

$$c_0(w, q, g, \Delta) = \mathrm{Var}[q'']^{-1}\mathrm{Cov}(q'', g')\mathrm{Var}[g']^{-1}\mathbb{E}\left[\frac{\partial g'}{\partial \lambda}\right]\left(\mathbb{E}\left[\frac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathbb{E}\left[\frac{\partial g'}{\partial \lambda}\right] + \Sigma_\Delta^{-1}\right)^{-1}$$
$$\cdot \mathbb{E}\left[\frac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathrm{Cov}(q'', g').$$

A.5 Linear Regression Experiment

We apply our calibrated noise scheme to a real-world linear regression dataset to highlight its improvement in limiting predictive exposure over an isotropic noise scheme. We use the California Housing dataset (Pace & Barry, 1997) with $N = 20,640$ and $d = 9$ (8 features and 1 additional bias feature). We compute the least squares regression vector $\hat{w}$ for the dataset,

$$\hat{w} = \arg\min_{w} \frac{1}{2}(\hat{w}^\top x - y)^2.$$

Let the query of interest $q(x)$ be the binary query such that $q(x) = 1$ if the median income is greater than 100K and $q(x) = 0$ otherwise. We set $w(x) = 1$; this models the setting where the attacker compromises a random subset of the dataset. We set the predictability threshold to be

$$\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) = 0.2 \cdot \frac{H(1 + \log(4))\mathrm{Var}[q(x)]}{n}.$$

This means that we would like to release a noisy regression vector $\tilde{w}$ such that only 20% of the variance of $q(x)$ is revealed through the release of $\tilde{w}$.

Let $\tilde{w}_1$ be the noisy regression vector using isotropic noise,

$$\tilde{w}_1 = \hat{w} + \mathcal{N}(0, \sigma_1^2 \cdot I_d).$$

Let $\tilde{w}_2$ be the noisy regression vector using tilted calibrated noise,

$$\tilde{w}_2 = \hat{w} + \mathcal{N}(0, \sigma_2^2 \cdot \mathbb{E}[xx^\top]^{-1}\mathbb{E}[(\hat{w}^\top x - y)^2 xx^\top]\mathbb{E}[xx^\top]^{-1}).$$

The noise scales σ_1 and σ_2 are selected to ensure the predictability threshold is satisfied. We plug in $\Sigma_\Delta = \sigma_1^2 \cdot I_d$ and $\Sigma_\Delta = \sigma_2^2 \cdot \mathbb{E}[xx^\top]^{-1}\mathbb{E}[(\hat{w}^\top x - y)^2 xx^\top]\mathbb{E}[xx^\top]^{-1}$, respectively, into the final expression of Corollary 1, set $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) = 0.2 \cdot \frac{H(1+\log(4))\mathrm{Var}[q(x)]}{n}$, and solve for σ_1 and σ_2.

Then we compare the average distortion in predictions between $\tilde{w}_1$ and $\tilde{w}_2$ with respect to $\hat{w}$. The average change in prediction when using

- Isotropic noise:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}_1^\top x_i - \hat{w}^\top x_i)^2\right] \approx 5,749,704$$

- Calibrated noise:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}_2^\top x_i - \hat{w}^\top x_i)^2\right] \approx 11,654$$

On average, the predictions of $\tilde{w}_2$ on the dataset are much closer to the predictions of the original ERM $\hat{w}$. The calibrated noise scheme leads to significantly less distortion of the privatized regression vector while achieving the same predictive exposure. Thus, the calibrated noise scheme has greater utility for the same predictive exposure. All of the above quantities can be efficiently computed given the dataset.

A.6 Details of TOFU Experiment

Noise Scales Let w be the model parameters of the fine-tuned Llama-3.2-1B-Instruct model. We add noise according to $w + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I)$ for isotropic noise and $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{E}[H]^{-1}\mathbb{E}[\nabla\nabla^\top]\mathbb{E}[H]^{-1})$ for calibrated noise. We used a single noise draw from $\mathcal{N}(0, I)$ and $\mathcal{N}(0, \mathbb{E}[H]^{-1}\mathbb{E}[\nabla\nabla^\top]\mathbb{E}[H]^{-1})$. Then these noise vectors are held fixed and scaled to different magnitudes for varying σ^2s. We tested 12 noise scales across a logarithmic scale. For isotropic noise, σ^2 ranged from 3×10^{-7} to 3×10^{-4}. For calibrated noise, σ^2 ranged from 0.01 to 2. These noise scales were selected so that the set of models would cover the same utility accuracy range. The noise draw from $\xi \sim \mathcal{N}(0, \mathbb{E}[H]^{-1}\mathbb{E}[\nabla\nabla^\top]\mathbb{E}[H]^{-1})$ was approximated using a matrix-free linear solve. Concretely, we first formed $b = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_i \nabla_i$ where z_i is drawn i.i.d. from $\mathcal{N}(0, 1)$ and ∇_i is the gradient of the i-th QA pair in the TOFU dataset evaluated at the model parameters. Thus, we gave $b \sim \mathcal{N}(0, \mathbb{E}[\nabla\nabla^\top])$. Then we approximately computed $\mathbb{E}[H]^{-1}\varepsilon$ using the minimal residual method (MINRES) (Paige & Saunders, 1975) with Hessian-vector products without explicitly computing the Hessian inverse.

Author Embeddings For each noisy model, we extracted the 2,048-dimensional output from the last (16th) transformer block (before the final RMS normalization) at the final token of each QA pair. Then we constructed one embedding per author by averaging together the representations from the author's corresponding 20 QA pairs.

Gender Classifier Head For each set of embeddings, we trained a small multilayer perceptron (MLP) to predict whether each author was female based on their 2,048-dimensional author embedding. The classifier consisted of a fully connected hidden layer with 16 units and ReLU activation. Input features were standardized using means and standard deviations estimated from the training authors. We optimized binary cross-entropy with an L2 regularization coefficient of 1.0 using Adam, a learning rate of 0.001, and a batch size of 32. Training ran for at most 2,000 epochs, with an early-stopping criterion of no training-loss improvement of at least 10^{-5} for 30 consecutive epochs. Predictions used a probability threshold of 0.5. For each set of embeddings, we repeated classifier training with three random seeds (42, 43, and 44), each specifying the classifier initialization and a random split of 100 training and 100 test authors. The same split was shared across all models within each seed. The shaded error bars in Figure 1 indicate ± 1 standard deviation across three classifier training seeds.

Measuring Model Utility For each noise level, we trained a separate next-token prediction head on frozen representations from the noisy fine-tuned model. The head received the 2,048-dimensional output of the last transformer block, before the model's final RMS normalization. The trained head is a residual bottleneck adapter with architecture $2048 \to 256 \to 2048$, using non-affine LayerNorm before the bottleneck and GELU activation between layers. The resulting representation from this head was then passed through the original fine-tuned model's frozen RMS normalization and vocabulary projection to obtain next-token logits. We trained the prediction head on all 4,000 QA pairs. We optimized the cross-entropy loss averaged first over answer tokens within each example and then over examples. Each head was trained for 20 epochs using AdamW with learning rate 0.001, weight decay 0.01, and gradient-norm clipping at 1.0. We evaluated utility on 217 multiple-choice questions from TOFU's real-author and world-fact datasets. Utility accuracy is the fraction of correctly selected answers.

A.7 Proofs from Section 2

A.7.1 Proposition 1

Proposition 1. *If $q : \mathcal{X} \to [K]$ is a categorical query and $\ell(\hat{\theta}, y) = -y_i \log(\hat{\theta}_i)$ is the log loss, then predictive exposure $\gamma(\mathcal{P}, q, \ell, A) = \sup_S \mathrm{MI}(q(x); A(S) \mid C(S))$.*

Proof. From the definition of predictive exposure, we know that $A(S)$ satisfies

$$\sup_S \mathbb{E}_{C(S)\sim\mathcal{P},A}[\mathbb{E}_{x\sim\Pi}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]] \le \gamma.$$

Under the log loss, the optimal $\hat{\theta}^*_{q|C(S)}$ and $\hat{\theta}^*_{q|C(S),A(S)}$ predict $\Pr[q(x) = i \mid C(S)]$ and $\Pr[q(x) = i \mid C(S), A(S)]$, respectively, for each $i \in [K]$. Plugging these in and expanding the loss,

$$\begin{aligned}
&\mathbb{E}_{x\sim\Pi,A}[\ell(\Pr[q(x) = i \mid C(S)], q(x)) - \ell(\Pr[q(x) = i \mid C(S), A(S)], q(x))] \\
&= \mathbb{E}_{x\sim\Pi,A}\left[-\sum_{i=1}^{K} \mathbb{1}\{y = i\} \log(\Pr[q(x) = i \mid C(S)]) + \sum_{i=1}^{K} \mathbb{1}\{y = i\} \log(\Pr[q(x) = i \mid C(S), A(S)])\right] \\
&= \mathbb{E}_{x\sim\Pi,A}\left[\sum_{i=1}^{K} \mathbb{1}\{y = i\} \log\left(\frac{\Pr[q(x) = i \mid C(S), A(S)]}{\Pr[q(x) = i \mid C(S)]}\right)\right] \\
&= \mathbb{E}_{x\sim\Pi,A}[D_{\mathrm{KL}}(P_{q(x)|C(S),A(S)} \| P_{q(x)|C(S)})].
\end{aligned}$$

Thus,

$$\sup_S \mathbb{E}_{C(S)\sim\mathcal{P}}[\mathbb{E}_{x\sim\Pi,A}[D_{\mathrm{KL}}(P_{q(x)|C(S),A(S)} \| P_{q(x)|C(S)})]] \le \gamma$$

From here, observe that

$$\sup_S \mathbb{E}_{C(S)\sim\mathcal{P}}[\mathbb{E}_{x\sim\Pi,A}[D_{\mathrm{KL}}(P_{q(x)|C(S),A(S)} \| P_{q(x)|C(S)})]] = \sup_S \mathrm{MI}(q(x); A(S) \mid C(S))$$

□

A.7.2 Proposition 2

Proposition 2. *Let $\ell(\hat{\theta}, y) = (\hat{\theta} - y)^2$ be the squared loss. There exists a process $\mathcal{P}$, query q, and algorithm A such that A is ε-differentially private, $\varepsilon = O(\frac{1}{\sqrt{N}})$, but $\gamma(\mathcal{P}, q, \ell, A)$-predictive exposure is $\Omega(1)$. Conversely, there exists a process $\mathcal{P}$, query q, and algorithm A such that $\gamma(\mathcal{P}, q, \ell, A)$-predictive exposure is $O(\frac{1}{N})$, but A is not differentially private.*

Proof. Let S be a binary dataset, $x_i \in \{0, 1\}$. Let $p = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the proportion of 1's. Consider a single query of interest $q(x) = x$.

For the first part of the statement, let $A(S)$ be an ε-differentially private release of p, $A(S) = p + \mathrm{Lap}(0, \frac{\varepsilon}{N})$, $\varepsilon = \frac{1}{\sqrt{N}}$. Consider a process $\mathcal{P}$ that only selects entries such that $x_i = 1$ to form $C(S)$ of size $\frac{N}{2}$. Let $p_{S\setminus C(S)}$ be the proportion of 1's. Given only $C(S)$, $\hat{\theta}^*_{q|C(S)}$ cannot estimate $p_{S\setminus C(S)}$ accurately. The best that $\hat{\theta}^*_{q|C(S)}$ can do is estimate 0.5 because $p_{S\setminus C(S)}$ can range anywhere from 0 to 1 and $C(S)$ is always a set of 1's.

However, given $C(S)$ and $A(S)$, $\hat{\theta}^*_{q|C(S),A(S)}$ can estimate $p_{S\setminus C(S)}$ very accurately. In particular,

$$\begin{aligned}
\frac{1}{2}p_{C(S)} + \frac{1}{2}p_{S\setminus C(S)} &= p \\
p_{S\setminus C(S)} &= 2p - p_{C(S)}
\end{aligned}$$

Consider when $\hat{\theta}^*_{q|C(S),A(S)}$ constructs the estimator

$$\begin{aligned}
\hat{p}_{S\setminus C(S)} &= 2 \cdot A(S) - p_{C(S)} \\
&= 2 \cdot (p + \Delta) - p_{C(S)} && (\Delta \sim \mathrm{Lap}(0, \tfrac{1}{\varepsilon N})) \\
&= p_{S\setminus C(S)} + 2\Delta \\
&= p_{S\setminus C(S)} + \Delta' && (\Delta \sim \mathrm{Lap}(0, \tfrac{2}{\varepsilon N}))
\end{aligned}$$

Let $\ell(\hat{y}, y) = (\hat{y} - y)^2$ be the squared loss. Consider the predictive exposure of q and plug in the estimators $\hat{\theta}^*_{q|C(S)}$ and $\hat{\theta}^*_{q|C(S),A(S)}$,

$$\begin{aligned}
&\sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]\Big] \\
&\quad= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A}[(\hat{\theta}^*_{q|C(S)} - q(x))^2 - (\hat{\theta}^*_{q|C(S),A(S)} - q(x))^2]\Big] \\
&\quad= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A}[(\hat{\theta}^*_{q|C(S)} - \mathop{\mathbb{E}}_{x\sim\Pi}[q(x)])^2 + \mathop{\mathbb{E}}_{x\sim\Pi}[(\mathop{\mathbb{E}}_{x\sim\Pi}[q(x)] - q(x))^2] - (\hat{\theta}^*_{q|C(S),A(S)} - \mathop{\mathbb{E}}_{x\sim\Pi}[q(x)])^2 \\
&\qquad - \mathop{\mathbb{E}}_{x\sim\Pi}[(\mathop{\mathbb{E}}_{x\sim\Pi}[q(x)] - q(x))^2]] \\
&\qquad\qquad\qquad

$$
\begin{aligned}
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[(p_{C(S)} - \mathop{\mathbb{E}}_{x\sim\Pi}[q(x)])^2 - (p_{S\setminus C(S))} - \mathop{\mathbb{E}}_{x\sim\Pi}[q(x)])^2] \\
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[(p_{C(S)} - p_{S\setminus C(S)})^2 - (p_{S\setminus C(S))} - p_{S\setminus C(S)})^2] \\
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[(p_{C(S)} - p_{S\setminus C(S)})^2] \\
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[((p_{C(S)} - p_S) - (p_{S\setminus C(S)} - p_S))^2] \\
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[((p_{C(S)} - p_S))^2] + \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[((p_{S\setminus C(S)} - p_S))^2] - 2\mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[(p_{C(S)} - p_S)(p_{S\setminus C(S)} - p_S)] \\
&\le O(\tfrac{1}{N}) + O(\tfrac{1}{N}) + 0 \\
&= O(\tfrac{1}{N}),
\end{aligned}
$$

where the last line follows from the concentration of $p_{C(S)}$ and $p_{S\setminus C(S)}$ around p_S. □

A.7.3 Proposition 3, Connection to Mutual Information Differential Privacy

Definition 13 (ε-mutual-information differential privacy (Cuff & Yu, 2016))**.** *A randomized algorithm $A(S)$ is ε-mutual-information differentially private (MI-DP) if*

$$\sup_{\chi\in\Delta(\mathcal{X}^N), i\in[N]} \mathrm{MI}(x_i; A(S) \mid S\setminus\{x_i\}) \le \varepsilon.$$

Proposition 3. *Let $\mathcal{P}_{\mathrm{DP}}$ be a process that reveals an arbitrary subset of $n = N-1$ individuals from the dataset S. Let $\mathcal{Q}_{\mathrm{bin}} = \{q : \mathcal{X} \mapsto \{0,1\}\}$ be the set of all binary queries on $\mathcal{X}$. Let ℓ be the log loss. If algorithm A has $\gamma(\mathcal{P}_{\mathrm{DP}}, \mathcal{Q}_{\mathrm{bin}}, \ell, A)$-predictive exposure γ, then A is γ-MI-DP.*

Proof. We know that $A(S)$ satisfies

$$\sup_S \sup_{q\in\mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_{C(S)\sim\mathcal{P}_{\mathrm{DP}},A}[\mathbb{E}_{x\sim\mathrm{Unif}(S\setminus C(S))}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]] \le \gamma.$$

Plugging in the form of $\mathcal{P}_{\mathrm{DP}}$, we have

$$\sup_S \sup_{q\in\mathcal{Q}_{\mathrm{bin}}} \sup_{i\in[N]} \mathbb{E}_A[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))]] \le \gamma.$$

This is equivalent to

$$\sup_{\chi\in\Delta(\mathcal{X}^N)} \mathbb{E}_{S\sim\chi}[\sup_{i\in[N]} \sup_{q\in\mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_A[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))]] \le \gamma.$$

Observe that

$$
\begin{aligned}
&\mathbb{E}_{S\sim\chi}[\sup_{i\in[N],q\in\mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_A[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))]] \\
&\qquad \ge \sup_{i\in[N],q\in\mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_{S\sim\chi,A}[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))]
\end{aligned}
$$

Thus,

$$\sup_{i\in[N],q\in\mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_{S\sim\chi,A}[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))] \le \gamma.$$

Under the log loss, the optimal $\hat{\theta}^*_{q|C(S)}$ and $\hat{\theta}^*_{q|C(S),A(S)}$ predict $\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}]$ and $\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}, A(S)]$, respectively.

Plugging these in and expanding the loss,

$$
\begin{aligned}
&\mathbb{E}_{S\sim\chi,A}[\ell(\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}], q(x_i)) - \ell(\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}, A(S)], q(x_i))] \\
&\quad = \mathbb{E}_{S\setminus\{x_i\}\sim\chi}[\mathbb{E}_{x_i|S\setminus\{x_i\},A}[\ell(\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}], q(x_i)) - \ell(\Pr[q(x_i) = 1 \mid S\setminus\{x_i\}, A(S)], q(x_i))]] \\
&\quad = \mathbb{E}_{S\setminus\{x_i\}\sim\chi}[D_{\mathrm{KL}}(P_{q(x)|S\setminus\{x_i\},A(S)} \| P_{q(x)|S\setminus\{x_i\}})] \qquad \text{(Proposition 1)}
\end{aligned}
$$

$$= \mathrm{MI}(q(x_i); A(S) \mid S \setminus \{x_i\})$$

where MI is the conditional mutual information between $q(x_i)$ and $A(S)$ given $S \setminus \{x_i\}$.

Thus, we have

$$\sup_{\chi \in \Delta(\mathcal{X}^N)} \sup_{i \in [N]} \sup_{q \in \mathcal{Q}_{\mathrm{bin}}} \mathrm{MI}(q(x_i); A(S) \mid S \setminus \{x_i\}) \leq \gamma,$$

which implies

$$\sup_{\chi \in \Delta(\mathcal{X}^N)} \sup_{i \in [N]} \mathrm{MI}(x_i; A(S) \mid S \setminus \{x_i\}) \leq \gamma.$$

Thus, $A(S)$ is γ-mutual-information differentially private (Definition 13). □

Theorem 13. *If A is ε-differentially private, then for process $\mathcal{P}_{\mathrm{DP}}$, A satisfies ε-predictive exposure for all binary queries $\mathcal{Q}_{\mathrm{bin}}$ under the log loss.*

Proof. Plugging in the form of $\mathcal{P}_{\mathrm{DP}}$ into predictive exposure, we look to bound

$$\sup_{S} \sup_{i \in [N]} \sup_{q \in \mathcal{Q}_{\mathrm{bin}}} \mathbb{E}_A[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))]$$

Since A is ε-differentially private, we know that for all S and i,

$$e^{-\varepsilon} \leq \frac{\Pr[A(S) \mid S]}{\Pr[A(S) \mid S \setminus \{x_i\}]} \leq e^{\varepsilon}.$$

Applying Bayes Rule, we know

$$\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)] = \frac{\Pr[A(S) \mid S \setminus \{x_i\}, q(x_i) = 1] \cdot \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]}{\Pr[A(S) \mid S \setminus \{x_i\}]}.$$

Let $\mathcal{S} = \{S' \mid S' \setminus \{x'_i\} = S \setminus \{x_i\}$ and $q(x'_i) = 1$. Then

$$\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)] = \frac{\sum_{S' \in \mathcal{S}} \Pr[A(S) \mid S'] \cdot \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]}{\Pr[A(S) \mid S \setminus \{x_i\}]}.$$

Thus, plugging in the ratio from differential privacy, we have

$$e^{-\varepsilon} \cdot \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}] \leq \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)] \leq e^{\varepsilon} \cdot \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}].$$

We can similarly conclude

$$e^{-\varepsilon} \cdot \Pr[q(x_i) = 0 \mid S \setminus \{x_i\}] \leq \Pr[q(x_i) = 0 \mid S \setminus \{x_i\}, A(S)] \leq e^{\varepsilon} \cdot \Pr[q(x_i) = 0 \mid S \setminus \{x_i\}].$$

In other words,

$$-\varepsilon \leq \log\left(\frac{\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)]}{\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]}\right), \log\left(\frac{\Pr[q(x_i) = 0 \mid S \setminus \{x_i\}, A(S)]}{\Pr[q(x_i) = 0 \mid S \setminus \{x_i\}]}\right) \leq \varepsilon.$$

Expanding out the log loss in the predictive exposure bound,

$$\begin{aligned}
&\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i)) \\
&\quad = \ell(\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}], q(x_i)) - \ell(\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)], q(x_i)) \\
&\quad = -\Pr[q(x_i) = 1] \cdot \log(\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]) - (1 - \Pr[q(x_i) = 1]) \cdot \log(1 - \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]) \\
&\qquad + \Pr[q(x_i) = 1] \cdot \log(\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)]) \\
&\qquad - (1 - \Pr[q(x_i) = 1]) \cdot \log(1 - \Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)]) \\
&\quad = \Pr[q(x_i) = 1] \cdot \log\left(\frac{\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}, A(S)]}{\Pr[q(x_i) = 1 \mid S \setminus \{x_i\}]}\right) + \Pr[q(x_i) = 0] \cdot \log\left(\frac{\Pr[q(x_i) = 0 \mid S \setminus \{x_i\}, A(S)]}{\Pr[q(x_i) = 0 \mid S \setminus \{x_i\}]}\right)
\end{aligned}$$

Since both

$$-\varepsilon \le \log\left(\frac{\Pr[q(x_i)=1 \mid S\setminus\{x_i\}, A(S)]}{\Pr[q(x_i)=1 \mid S\setminus\{x_i\}]}\right), \log\left(\frac{\Pr[q(x_i)=0 \mid S\setminus\{x_i\}, A(S)]}{\Pr[q(x_i)=0 \mid S\setminus\{x_i\}]}\right) \le \varepsilon,$$

we must have

$$-\varepsilon \le \ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i)) \le \varepsilon.$$

The above holds for every S, i, and q; thus, we can conclude

$$\sup_S \sup_{i\in[N]} \sup_{q\in\mathcal{Q}_{\text{bin}}} \mathbb{E}_A[\ell(\hat{\theta}^*_{q|S\setminus\{x_i\}}, q(x_i)) - \ell(\hat{\theta}^*_{q|S\setminus\{x_i\},A(S)}, q(x_i))] \le \gamma.$$

□

A.8 Post-Processing and Composition of Predictive Exposure

Theorem 14 (Post-Processing)**.** *If $A(S)$ has $\gamma(\mathcal{P},\mathcal{Q},\ell,A)$-predictive exposure γ, then $f\circ A$ has $\gamma(\mathcal{P},\mathcal{Q},\ell,f\circ A))$-predictive exposure γ.*

Proof. Notice that due to the monotonicity of Bayes risk, the Bayes optimal predictor must satisfy

$$\mathbb{E}[\ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))] \le \mathbb{E}[\ell(\hat{\theta}^*_{q|C(S),f(A(S))}, q(x))].$$

Thus,

$$\sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A}\Big[\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),f(A(S))}, q(x))]\Big]$$
$$\le \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A}\Big[\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]\Big]$$

□

Theorem 15 (Composition)**.** *For two algorithms A_1 and A_2, the $\gamma(\mathcal{P},\mathcal{Q},\ell,A)$-predictive exposure of the composed $(A_1(S),A_2(S))$ is bounded by*

$$\gamma(\mathcal{P},\mathcal{Q},\ell,(A_1,A_2)) \le \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A_1}\Big[\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x))]\Big]$$
$$+ \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P},A_1,A_2}\Big[\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S),A_2(S)}, q(x))]\Big]$$

Proof.

$$\sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A_1,A_2}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S),A_2(S)}, q(x))]\Big]$$
$$= \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A_1,A_2}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x))$$
$$+ \ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S),A_2(S)}, q(x))]\Big]$$
$$\le \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A_1}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x))]\Big]$$
$$+ \sup_S \sup_{q\in\mathcal{Q}} \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi,A_1,A_2}[\ell(\hat{\theta}^*_{q|C(S),A_1(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A_1(S),A_2(S)}, q(x))]\Big]$$

where the last line follows from the subadditivity of supremum. □

A.9 Proofs from Section 4

A.9.1 Lemma 1

Lemma 1. *Let $\ell(\hat{\theta}) = \mathbb{E}_{x\sim\Pi}[\ell(\hat{\theta}, q(x)]$ be a convex, H-smooth loss. Let $\theta_0^q = \mathbb{E}_{x\sim\Pi}[q(x)]$. Then for any pair of estimators $\hat{\theta}_{q|C(S)}$ and $\hat{\theta}_{q|C(S),A(S)}$,*

$$\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}_{q|C(S)}, q(x)) - \ell(\hat{\theta}_{q|C(S),A(S)}, q(x))] \le H|\hat{\theta}_{q|C(S)} - \theta_0^q| \cdot |\hat{\theta}_{q|C(S)} - \hat{\theta}_{q|C(S),A(S)}|.$$

For log loss, if the true $p_0 = \Pr[q(x)=1]$ satisfies $p_0 \in [\tau, 1-\tau]$, $H = \frac{1}{\tau^2}$. For squared loss, $H = 2$.

Proof.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i\sim\Pi}[\ell(\hat{\theta}_{q|C(S)}, q(x)) - \ell(\hat{\theta}_{q|C(S),A(S)}, q(x))] \\
&\qquad = \ell(\hat{\theta}_{q|C(S)}) - \ell(\hat{\theta}_{q|C(S),A(S)}) \\
&\qquad \le \ell'(\hat{\theta}_{q|C(S)})|\hat{\theta}_{q|C(S)} - \hat{\theta}_{q|C(S),A(S)}| && \text{(convexity)} \\
&\qquad \le H|\hat{\theta}_{q|C(S)} - \theta_0^q| \cdot |\hat{\theta}_{q|C(S)} - \hat{\theta}_{q|C(S),A(S)}| && \text{(smoothness)}
\end{aligned}$$

□

A.9.2 Regularity Conditions

Define sample quantities

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, \theta)$$

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_i, \theta)$$

$$\Omega_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, \theta) f(x_i, \theta)^\top$$

and define population quantities

$$\begin{aligned}
\mathbb{E}[f(\theta)] &= \mathbb{E}[f(x, \theta)] \\
G(\theta) &= \mathbb{E}[\nabla f(x, \theta)] \\
\Omega(\theta) &= \sum_{k=-\infty}^{\infty} f(x_0, \theta) f(x_k, \theta)^\top.
\end{aligned}$$

Assume the following regularity conditions.

- The domain Θ of θ is a compact subset of $\mathbb{R}^r$ and θ_0 is its interior.
- $f(x, \theta)$ is $\mathbb{R}^r \mapsto \mathbb{R}^m$ function of θ, where $r \le m$.
- f is continuous in θ and continuously differentiable on a neighborhood of θ_0 in θ.
- $\mathbb{E}[f(\theta)] = 0$ if and only if $\theta = \theta_0$.
- $G(\theta_0)$ is a $m \times r$ matrix of rank r.
- $\Omega(\theta_0)$ is a $m \times m$ positive definite matrix.
- W is a $m \times m$ positive definite matrix.
- $f(x, \theta)$, $G(\theta)$, and $\Sigma(\theta)$ are dominated (there exists a function $\beta(x)$ with $\mathbb{E}[\beta(x)] < +\infty$ such that $|f(x, \theta)| \le \beta(x)$, $|G(\theta)| \le \beta(x)$ and $|\Omega(\theta)| \le \beta(x)$).
- For each θ, $f_n(\theta)$ obeys a central limit theorem (Theorem 17)
- $f_n(\theta)$, $G_n(\theta)$, and $\Omega_n(\theta)$ obey a uniform law of large numbers (Theorem 16).

Theorem 16 (Uniform Law of Large Numbers for Stationary Ergodic Processes (Hansen, 1982))**.** *Let $Y_1(\theta), \ldots, Y_n(\theta)$ be from a stationary ergodic process. Define*

$$\psi(\theta) = \mathbb{E}[Y_1(\theta)], \quad \bar{Y}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\theta).$$

Assume a finite mean $\psi(\theta) < +\infty$ for all θ in a closed bounded set $\Theta \subseteq \mathbb{R}^k$. Assume $Y_i(\theta)$ is continuous at each $\theta \in \Theta$. Assume that $Y_i(\theta)$ is dominated; there exists a random variable Z with a finite mean such that $\sup_{\theta\in\Theta} |Y_i(\theta)| \le Z$. Then

$$\sup_{\theta\in\Theta} |\bar{Y}_n(\theta) - \psi(\theta)| \to_p 0$$

Theorem 17 (Central Limit Theorem for Stationary Ergodic Processes (Hansen, 1982))**.** *Let $X_1, \ldots, X_n$ be from a stationary ergodic process. Assume (X_t) is α-mixing with mixing coefficients satisfying*

$$\sum_{k=1}^{\infty} \alpha(k)^{\delta/(2+\delta)} < \infty.$$

As $n \to \infty$,

$$\sqrt{n}(\bar{X}_n - \theta) \to_d \mathcal{N}(0, \Omega_\infty)$$

where θ is the population mean and $\Omega_\infty = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(X_0, X_k)$ is the long-run covariance.

A.9.3 THEOREM 1, GMM EFFICIENCY

Theorem 1 (Efficient GMM Estimator Hansen (1982)). *Let* $G = \mathbb{E}[\frac{\partial f(x,\theta)}{\partial \theta}]$ *be the Jacobian and let* $\Omega = \sum_{k=-\infty}^{\infty} \mathbb{E}[f(x_0,\theta)f(x_k,\theta)^\top]$ *be the long-run covariance. As* $n \to \infty$*, with* $W = \Omega^{-1}(\theta_0)$*, the efficient GMM estimator* $\hat{\theta}^{\text{GMM}}$ *is regular and asymptotically efficient with asymptotic distribution*

$$\sqrt{n}(\hat{\theta}^{\text{GMM}} - \theta_0) \sim \mathcal{N}(0, [G(\theta_0)^\top \Omega^{-1}(\theta_0) G(\theta_0)]^{-1}).$$

This is a standard result from Hansen (1982).

For completeness, we include a proof of the efficient GMM estimator for sampling without replacement. In particular, consider sample quantities

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta)$$
$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_i,\theta)$$
$$\Omega_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta) f(x_i,\theta)^\top$$

and population quantities

$$\mathbb{E}[f(\theta)] = \frac{1}{N}\sum_{i=1}^{N} f(x_i,\theta)$$
$$G(\theta) = \frac{1}{N}\sum_{i=1}^{N} \nabla f(x_i,\theta)$$
$$\Omega(\theta) = \frac{1}{N}\sum_{i=1}^{N} f(x_i,\theta) f(x_i,\theta)^\top.$$

Consider the specific form of ULLN and CLT for sampling without replacement.

Theorem 18 (Uniform Law of Large Numbers for Sampling without Replacement (Horowitz, 1990)). *Let* $Y_1(\theta),\dots,Y_n(\theta)$ *be a simple random sample without replacement from finite population* $\{Y_1(\theta),\dots,Y_N(\theta)\}$*, where* $n/N = \nu$*,* $0 < \nu < 1$*. Define*

$$\psi(\theta) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\theta), \quad \bar{Y}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\theta).$$

Assume a finite mean $\psi(\theta) < +\infty$ *for all* θ *in a closed bounded set* $\Theta \subseteq \mathbb{R}^k$*. Assume* $Y_i(\theta)$ *is continuous at each* $\theta \in \Theta$*. Assume that* $Y_i(\theta)$ *is dominated; there exists a random variable* Z *with a finite mean such that* $\sup_{\theta\in\Theta} |Y_i(\theta)| \le Z$*. Then*

$$\sup_{\theta\in\Theta} |\bar{Y}_n(\theta) - \psi(\theta)| \to_p 0$$

Theorem 19 (Central Limit Theorem for Sampling without Replacement (Prášková & Sen, 2009)). *When sampling* n *elements from* N *elements uniformly at random, where* $n/N = \nu$*,* $0 < \nu < 1$*. As* $n, N \to \infty$*,*

$$\sqrt{\frac{n}{1-\nu}}(\bar{X}_n - \theta) \to_d \mathcal{N}(0, F)$$

where θ, F *are the population mean and covariance matrix.*

Theorem 20 (Optimal GMM Estimator for Sampling without Replacement). *Consider a sample without replacement* $x_1,\dots,x_n$ *from a larger population of* N *elements, where* $n/N = \nu$ *and* $0 < \nu < 1$ *is a constant. Given the sample, the GMM estimator of* θ *is defined as*

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta\subseteq\mathbb{R}^r} f_n(\theta)^\top W f_n(\theta).$$

Assume the regularity conditions from Appendix A.9.2 hold. Then $\hat{\theta}_n$ is asymptotically normal,

$$\sqrt{\frac{n}{1-\nu}}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}(0, (G(\theta_0)^\top W G(\theta_0))^{-1} G(\theta_0)^\top W \Omega(\theta_0) W G(\theta_0) (G(\theta_0)^\top W G(\theta_0))^{-1}).$$

With optimal choice $W = \Omega(\theta_0)$, $\hat{\theta}_n^{\text{GMM}}$ is asymptotically normal as

$$\sqrt{\frac{n}{1-\nu}}(\hat{\theta}_n^{\text{GMM}} - \theta_0) \to_d \mathcal{N}(0, (G(\theta_0)^\top W G(\theta_0))^{-1}.$$

and is asymptotically efficient among the class of all regular estimators that only use the information contained in the moment conditions.

Proof. The proof follows similarly to the standard GMM proof for IID samples from McFadden (1999), Chapter 3, Theorem 3.1.

Let $Q_n = f_n(\theta)^\top W f_n(\theta)$.

For every fixed θ, $f_n(\theta)$, $G_n(\theta)$, and $\Omega_n(\theta)$ converge in probability to $\mathbb{E}[f(x,\theta)]$, $G(\theta)$, and $\Omega(\theta)$, respectively, using Theorem 18. Furthermore,

$$\Omega(\theta_0)^{-1/2} \sqrt{\frac{n}{1-\nu}}(f_n(\theta_0)) \equiv U_n \to_d U \sim \mathcal{N}(0, I).$$

For each fixed θ, the uniform law of large numbers implies that $f_n(\theta) \to_p \mathbb{E}[f(\theta)]$. This implies that $Q_n(\theta_0) \to_p \mathbb{E}[f(\theta_0)]^\top W \mathbb{E}[f(\theta_0)] = 0$. Thus, $\frac{n}{1-\nu} \cdot Q_n(\theta_0)$ is stochastically bounded ($\frac{n}{1-\nu} \cdot Q_n(\theta_0)$ does not go to infinity in probability).

Consider any estimator T_n^* that satisfies $Q_n(T_n^*) \to_p 0$. For each fixed θ, Theorem 18 implies that $f_n(\theta) \to_p \mathbb{E}[f(\theta)]$. Then applying Theorem 27, $Q_n(\theta) \to_p \mathbb{E}[f(\theta)]^\top W \mathbb{E}[f(\theta)]$. Since $\mathbb{E}[f(\theta)] = 0$ only when $\theta = \theta_0$, then $T_n^* \to_p \theta_0$. T_n^* is consistent.

Consider any estimator T_n^* that satisfies $\frac{n}{1-\nu} \cdot Q_n(T_n^*)$ is stochastically bounded. This implies that $Q_n(T_n^*) \to_p 0$, which means $T_n^* \to_p \theta_0$. The mean value theorem and the CLT condition for $f_n(\theta)$ give

$$\sqrt{\frac{n}{1-\nu}}(f_n(T_n^*)) = \sqrt{\frac{n}{1-\nu}}(f_n(\theta_0)) - G_n(T_n^{\text{MID}}) \sqrt{\frac{n}{1-\nu}}(T_n^* - \theta_0)$$
$$\sqrt{\frac{n}{1-\nu}}(f_n(T_n^*)) = \Omega(\theta_0)^{1/2} U_n - G_n(T_n^{\text{MID}}) \sqrt{\frac{n}{1-\nu}}(T_n^* - \theta_0)$$

for some intermediate value T_n^{MID}. Then using the triangle inequality, $\|a-b\|_W^2 \le \|a\|_W^2 + \|b\|_W^2$,

$$\left\| G_n(T_n^{\text{MID}}) \sqrt{\frac{n}{1-\nu}}(T_n^* - \theta_0) \right\|_W^2 \le \left\| \sqrt{\frac{n}{1-n/N}}(f_n(T_n^*)) \right\|_W^2 + \left\| \Omega(\theta_0)^{1/2} U_n \right\|_W^2$$

$$\frac{n}{1-\nu} \cdot (T_n^* - \theta_0) G_n(T_n^{\text{MID}})^\top W G_n(T_n^{\text{MID}}) (T_n^* - \theta_0) \le \frac{n}{1-\nu} \cdot Q_n(T_n^*) + \frac{1}{2} U_n^\top \Omega(\theta_0)^{1/2} W \Omega(\theta_0)^{1/2} U_n$$

U_n and $\frac{n}{1-\nu} \cdot Q_n(T_n^*)$ are stochastically bounded, so $\frac{n}{1-\nu} \cdot (T_n^* - \theta_0) G_n(T_n^{\text{MID}})^\top W G_n(T_n^{\text{MID}}) (T_n^* - \theta_0)$ is stochastically bounded.

Since we have $T_n^{\text{MID}} \to_p \theta_0$ and uniform convergence of $G_n(\theta)$, Theorem 28 implies that $G_n(T_n^{\text{MID}})^\top W G_n(T_n^{\text{MID}}) \to_p G(\theta_0)^\top W G(\theta_0)$, where $G(\theta_0)^\top W G(\theta_0)$ is positive definite. Let λ be the smallest eigenvalue of $G(\theta)^\top W G(\theta)$. Then in the probability limit,

$$\lambda \cdot \left\| \sqrt{\frac{n}{1-\nu}}(T_n^* - \theta_0) \right\|^2 \le \frac{n}{1-\nu} \cdot (T_n^* - \theta_0) G_n(T_n^{\text{MID}})^\top W G_n(T_n^{\text{MID}}) (T_n^* - \theta_0).$$

Thus, $\sqrt{\frac{n}{1-\nu}} \cdot (T_n^* - \theta_0)$ must be stochastically bounded. This in turn implies that $\sqrt{\frac{n}{1-\nu}}(f_n(T_n^*))$ is stochastically bounded.

Consider the estimator $T_n = \arg\min Q_n(\theta)$. We must have $Q_n(T_n) \le Q_n(\theta_0)$. $Q_n(\theta_0)$ and $Q_n(T_n)$ are stochastically bounded, which implies that T_n is consistent and $\sqrt{\frac{n}{1-\nu}}(f_n(T_n))$ is stochastically bounded, by applying the previous steps. T_n must satisfy the first order condition

$G_n(T_n)^\top W\sqrt{\frac{n}{1-\nu}} f_n(T_n) = 0$. Using the mean value theorem, the first order condition is equivalent to

$$-G_n(T_n)^\top W\Omega(\theta_0)^{1/2}U_n + G_n(T_n)^\top W G_n(T_n^{\text{MID}})\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0)$$

In the probability limit,

$$\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0) = (G_n(T_n)^\top W G_n(T_n^{\text{MID}}))^{-1} G_n(T_n)^\top W\Omega(\theta_0)^{1/2}U_n$$

exists because $G_n(T_n^{\text{MID}})^\top W G_n(T_n) \to_p G(\theta_0)^\top W G(\theta_0)$, and $G(\theta)^\top W G(\theta)$ is positive definite. Then in the limit,

$$\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0) \to_d (G(\theta_0)^\top W G(\theta_0))^{-1} G(\theta_0)^\top W\Omega(\theta_0)^{1/2}U$$

where $G_n(T_n^{\text{MID}})^\top W G_n(T_n)$ converges in probability and in distribution to $(G(\theta_0)^\top W G(\theta_0))^{-1}$, $G_n(T_n)^\top W\Omega(\theta_0)^{1/2}$ converges in probability and in distribution to $G(\theta_0)^\top W\Omega(\theta_0)^{1/2}$, and U_n converges in distribution to U. Thus,

$$\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0) \to_d \mathcal{N}(0, (G(\theta_0)^\top W G(\theta_0))^{-1} G(\theta_0)^\top W\Omega(\theta_0) W G(\theta_0)(G(\theta_0)^\top W G(\theta_0))^{-1}).$$

The choice of $W = \Omega(\theta_0)$ leads to the smallest covariance of $(G(\theta_0)^\top\Omega(\theta_0)G(\theta_0))^{-1}$ using Theorem 26.

Finally, following Hansen (1982) and Proposition 3.6 from Dovonon et al. (2025), we justify why the choice of $W = \Omega(\theta_0)$ is asymptotically efficient.

Consider any consistent regular estimator that only uses the moment conditions for estimation. It must lie in the span of the moment conditions and take the form, $A^\top f_n$, and have asymptotic variance $A^\top\Omega A$, since $\mathbb{E}[f_n(\theta_0)f_n(\theta_0)^\top] = \Omega$. The consistent estimator must also satisfy the first-order condition $A^\top G = I$. Consider the minimum possible variance that can be achieved,

$$\min_{A \text{ s.t. } A^\top G = I} A^\top\Omega A = (G^\top\Omega G)^{-1},$$

thus demonstrating optimality and asymptotic efficiency among the class of all estimators that only use the moment conditions for any estimator that achieves asymptotic variance $(G^\top\Omega G)^{-1}$. □

Theorem 2. *Let $\ell(\hat\theta) = \mathbb{E}_{x\sim\Pi}[\ell(\hat\theta, q(x)]$ be a twice continuously differentiable loss function with $\nabla\ell(\theta_0) = 0$ and $\nabla^2\ell(\theta_0)$ positive definite. Then*

$$\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat\theta^*_{q|C(S)}, q(x)) - \ell(\hat\theta^*_{q|C(S),A(S)}, q(x))] \le \mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat\theta^{\text{GMM}}_{q|C(S)}, q(x)) - \ell(\hat\theta^{\text{GMM}}_{q|C(S),A(S)}, q(x))] + o(\tfrac{1}{n})$$

Proof.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i\sim\Pi}[\ell(\hat\theta^*_{q|C(S)}, q(x)) - \ell(\hat\theta^*_{q|C(S),A(S)}, q(x))] - \mathop{\mathbb{E}}_{x_i\sim\Pi}[\ell(\hat\theta^{\text{GMM}}_{q|C(S)}, q(x)) - \ell(\hat\theta^{\text{GMM}}_{q|C(S),A(S)}, q(x))] \\
&\quad= \ell(\hat\theta^*_{q|C(S)}) - \ell(\hat\theta^*_{q|C(S),A(S)}) - \ell(\hat\theta^{\text{GMM}}_{q|C(S)}) + \ell(\hat\theta^{\text{GMM}}_{q|C(S),A(S)}) \\
&\quad= \ell(\hat\theta^*_{q|C(S)}) - \ell(\hat\theta^{\text{GMM}}_{q|C(S)}) + \ell(\hat\theta^{\text{GMM}}_{q|C(S),A(S)}) - \ell(\hat\theta^*_{q|C(S),A(S)}) \\
&\quad\le o\left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right) \qquad \text{(Theorem 21)} \\
&\quad= o\left(\frac{1}{n}\right)
\end{aligned}$$

We can apply Theorem 21 because both $\hat\theta^*$ and $\hat\theta^{\text{GMM}}$ are asymptotically efficient, asymptotically normal, and regular. □

Theorem 21. *Let $\hat\theta_1$ and $\hat\theta_2$ be any two asymptotically efficient and asymptotically normal regular estimators with respect to the same observation and information. Let $\ell(\theta)$ be a twice continuously differentiable loss function with $\nabla\ell(\theta_0) = 0$ and $\nabla^2\ell(\theta_0)$ positive definite. Then $\ell(\hat\theta_1) - \ell(\hat\theta_2) = o(\frac{1}{n})$.*

Proof. Since both $\hat{\theta}_1$ and $\hat{\theta}_2$ are regular and asymptotically efficient, they admit the same influence function $\psi(X)$. Thus, we express each estimator as

$$\sqrt{n}(\hat{\theta}_1 - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o(1),$$
$$\sqrt{n}(\hat{\theta}_2 - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o(1).$$

Taking the difference of the above, we have

$$\sqrt{n}\big(\hat{\theta}_1 - \hat{\theta}_2\big) = o(1)$$
$$\hat{\theta}_1 - \hat{\theta}_2 = o\left(\frac{1}{\sqrt{n}}\right)$$

Next, following the delta method from Lemma 3, we perform a second-order Taylor expansion of $\ell(\hat{\theta})$ around θ_0.

$$\begin{aligned}\ell(\hat{\theta}) &= \ell(\theta_0) + \nabla\ell(\theta_0)^\top(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top \nabla^2\ell(\theta_0)(\hat{\theta} - \theta_0) + o\left(\|\hat{\theta} - \theta_0\|^2\right) \\ &= \ell(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top \nabla^2\ell(\theta_0)(\hat{\theta} - \theta_0) + o\left(\|\hat{\theta} - \theta_0\|^2\right) && (\nabla\ell(\theta_0) = 0)\end{aligned}$$

Then consider $\ell(\hat{\theta}_1) - \ell(\hat{\theta}_2)$, and plug in the second order expansion.

$$\begin{aligned}\ell(\hat{\theta}_1) - \ell(\hat{\theta}_2) &= \frac{1}{2}(\hat{\theta}_1 - \theta_0)^\top \nabla^2\ell(\theta_0)(\hat{\theta}_1 - \theta_0) - \frac{1}{2}(\hat{\theta}_2 - \theta_0)^\top \nabla^2\ell(\theta_0)(\hat{\theta}_2 - \theta_0) + o\left(\frac{1}{n}\right) \\ &= \frac{1}{2}(\hat{\theta}_1 - \hat{\theta}_2)^\top \nabla^2\ell(\theta_0)(\hat{\theta}_1 - \hat{\theta}_2) + o\left(\frac{1}{n}\right) \\ &= o\left(\frac{1}{n}\right) && (\hat{\theta}_1 - \hat{\theta}_2 = o(\tfrac{1}{\sqrt{n}}))\end{aligned}$$

□

Lemma 3 (Delta Method)**.** *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta_0) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and θ_0, suppose that $g'(\theta_0) = 0$ and $g''(\theta_0)$ exists and is nonzero. Then*

$$n(g(Y_n) - g(\theta_0)) \to \sigma^2 \frac{g''(\theta_0)}{2} \chi_1^2.$$

A.9.4 Theorem 3, Difference of GMM Estimators

Theorem 3. *When the efficient GMM estimators $\hat{\theta}^{\text{GMM}}_{q|C(S)}$ and $\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}$ are constructed on the same sample, their difference has asymptotic distribution*

$$\hat{\theta}^{\text{GMM}}_{q|C(S)} - \hat{\theta}^{\text{GMM}}_{q|C(S),A(S)} \sim \mathcal{N}(0, \text{Var}[\hat{\theta}^{\text{GMM}}_{q|C(S)}] - \text{Var}[\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}])$$

Proof. We know

$$\Omega(\theta_0)^{-1/2} \sqrt{\frac{n}{1-\nu}} (f_n(\theta_0)) \equiv U_n \to_d U \sim \mathcal{N}(0, I).$$

We have, in the limit, for a GMM predictor T_n

$$\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0) \to_d (G(\theta_0)^\top W G(\theta_0))^{-1} G(\theta_0)^\top W \Omega(\theta_0)^{1/2} U_n$$

Thus, T_n is asymptotically linear with

$$\sqrt{\frac{n}{1-\nu}}(T_n - \theta_0) \to_d (G^\top W G)^{-1} G^\top W \sqrt{\frac{n}{1-\nu}} (f_n(\theta_0))$$

Let $(G^\top WG))^{-1}G^\top W = A^\top$. This predictor has covariance

$$A^\top \mathbb{E}\left[\frac{n}{1-\nu} f_n(\theta_0) f_n(\theta_0)^\top\right] A.$$

All choices of A must satisfy $A^\top G = I$ for the first order condition. The optimal A also minimizes the above covariance.

$$A^* = \arg \min_{A \text{ s.t. } A^\top G = I} A^\top \Omega A$$

where $\mathbb{E}[f_n(\theta_0) f_n(\theta_0)^\top] = \Omega$.

Consider another $\bar{A}$ that satisfies $\bar{A}^\top G = I$. The optimality condition of A^* implies that $\Omega A^* = G\Lambda$ where Λ is the Lagrangian multiplier. Consider

$$\begin{aligned}(\bar{A} - A^*)^\top \Omega A &= (\bar{A} - A^*)^\top G\Lambda \\ &= (\bar{A}^\top G - A^{*\top} G)\Lambda \\ &= (I - I)\Lambda \\ &= 0\end{aligned}$$

The predictor that only uses the first half of the moment conditions in f_n is asymptotically linear with

$$\sqrt{\frac{n}{1-\nu}}(T_n^1 - \theta_0) \to_d (G_1^\top W_1 G_1)^{-1} G_1^\top W_1 \sqrt{\frac{n}{1-\nu}}(f_1(\theta_0))$$

Let $(G_1^\top W_1 G_1))^{-1} G_1^\top W_1 = A_1^\top$. A_1 satisfies $A_1^\top G_1 = I$ by the first order condition. Let $\bar{A}^\top = [A_1^\top \quad 0]$. Then

$$\begin{aligned}\bar{A}^\top G &= [A_1^\top \quad 0]\, G \\ &= [A_1^\top \quad 0]\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \\ &= A_1^\top G_1 \\ &= I\end{aligned}$$

Thus, we must have $(\bar{A} - A^*)^\top \Omega A = 0$. Thus, $\bar{A}^\top \Omega A^* = A^{*\top} \Omega A^*$. $\bar{A}^\top \Omega A^*$ is the covariance between $(T_n^1 - \theta_0)$ and $(T_n - \theta_0)$ (when T_n is the optimal predictor). $A^{*\top} \Omega A^*$ is the variance of $(T_n - \theta_0)$ (when T_n is the optimal predictor). Thus, $\mathrm{Cov}(T_n^1, T_n) = \mathrm{Var}[T_n]$, and $\mathrm{Var}[T_n^1 - T_n] = \mathrm{Var}[T_n^1] - \mathrm{Var}[T_n]$.

Thus, in the limit, $T_n^1 - T_n$ is distributed $\mathcal{N}(0, \mathrm{Var}[T_n^1] - \mathrm{Var}[T_n])$. Here, $\hat{\theta}^*_{q|C(S)}$ is exactly T_n^1, and $\hat{\theta}^*_{q|C(S),A(S)}$ is exactly T_n. □

A.9.5 Theorem 4, Asymptotic Predictive Exposure of a Query

Theorem 4 (Asymptotic Predictive Exposure)**.** *Let $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. Assume process $\mathcal{P}$ satisfies Assumption 1 and the moment conditions in Equation 1 satisfy the regularity conditions (Appendix A.9.2). Let $\sigma_1 = \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}]$ and $\sigma_2 = \mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}]$. As $n, N \to \infty$, $n/N \to 0$, for a convex, H-smooth loss, with probability $1 - \delta$,*

$$\mathop{\mathbb{E}}_{x\sim\Pi}[\ell(\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}, q(x)) - \ell(\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}, q(x))]] \le \frac{H\log(4/\delta)}{n} \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}.$$

Let $q' = w(x)(q(x) - p_0)$ and let $g' = w(x)g(x, \lambda_0)$. Let

$$\begin{aligned}c_0(w, q, g, \Delta) = \mathrm{Var}[q']^{-1}\mathrm{Cov}(q', g')\mathrm{Var}[g']^{-1}\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]\left(\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right] + \Sigma_\Delta^{-1}\right)^{-1} \\ \cdot \mathbb{E}\left[\tfrac{\partial g'}{\partial \lambda}\right]^\top \mathrm{Var}[g']^{-1}\mathrm{Cov}(q', g').\end{aligned}$$

Deriving the form of σ_1 and σ_2, we have that $A(S)$ has predictive exposure

$$\gamma(\mathcal{P}, q, \ell, A) \le \frac{H(1 + \log(4))}{n} \cdot \mathrm{Var}[q'] \cdot \sqrt{\rho_{\mathrm{cc}}(q', g')^2 - c_0(w, q, g, \Delta)}.$$

Proof.

$$\begin{aligned}
\gamma(\mathcal{P}, q, \ell, A) &= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi, A}[\ell(\hat{\theta}^*_{q|C(S)}, q(x)) - \ell(\hat{\theta}^*_{q|C(S),A(S)}, q(x))]\Big] \\
&\le \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{x\sim\Pi, A}[\ell(\hat{\theta}^{\text{GMM}}_{q|C(S)}, q(x)) - \ell(\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}, q(x))]\Big] + o\left(\frac{1}{n}\right) \\
&\qquad \text{(using Theorem 2)} \\
&\le \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[H \cdot |\hat{\theta}^{\text{GMM}}_{q|C(S)} - \theta^q_0| \cdot |\hat{\theta}^{\text{GMM}}_{q|C(S)} - \hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}|]] + o\left(\frac{1}{n}\right) \\
&\qquad \text{(using Lemma 1)} \\
&= \sup_S \mathop{\mathbb{E}}_{C(S)\sim\mathcal{P}}[H \cdot |\hat{\theta}^{\text{GMM}}_{q|C(S)} - \theta^q_0| \cdot |\hat{\theta}^{\text{GMM}}_{q|C(S)} - \hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}|]]
\end{aligned}$$

where asymptotically, we drop the lower-order term.

Then we bound $|\hat{\theta}^{\text{GMM}}_{q|C(S)} - \theta^q_0|$ and $|\hat{\theta}^{\text{GMM}}_{q|C(S)} - \hat{\theta}^{\text{GMM}}_{q|C(S),A(S)}|$ using the behavior of the estimators.

We know $\hat{\theta}^{\text{GMM}}_{q|C(S)} = \theta_1 \sim \mathcal{N}(\theta_0, \frac{\sigma_1^2}{n})$ and $\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)} = \theta_2 \sim \mathcal{N}(\theta_0, \frac{\sigma_2^2}{n})$. Thus,

$$\Pr[|\theta_1 - \theta_0| \ge t] \le 2e^{-t^2 n/\sigma_1^2}$$

We want this probability to be less than $\delta/2$.

$$\begin{aligned}
\frac{\delta}{2} &= 2e^{-t^2 n/\sigma_1^2} \\
\frac{\delta}{4} &= e^{-t^2 n/\sigma_1^2} \\
\log(4/\delta) &= \frac{t^2 n}{\sigma_1^2} \\
\frac{\sigma_1^2 \log(4/\delta)}{n} &= t^2 \\
\sqrt{\frac{\sigma_1^2 \log(4/\delta)}{n}} &= t
\end{aligned}$$

We know $\theta_1 - \theta_2 \sim \mathcal{N}(0, \frac{\sigma_1^2}{n} - \frac{\sigma_2^2}{n})$.

$$\Pr[|\theta_1 - \theta_2| \ge t] \le 2e^{-t^2 n/(\sigma_1^2 - \sigma_2^2)}$$

We want this probability to be less than $\delta/2$.

$$\begin{aligned}
\frac{\delta}{2} &= 2e^{-t^2 n/(\sigma_1^2 - \sigma_2^2)} \\
\frac{\delta}{4} &= e^{-t^2 n/(\sigma_1^2 - \sigma_2^2)} \\
\log(4/\delta) &= \frac{t^2 n}{(\sigma_1^2 - \sigma_2^2)} \\
\frac{(\sigma_1^2 - \sigma_2^2)\log(4/\delta)}{n} &= t^2 \\
\sqrt{\frac{(\sigma_1^2 - \sigma_2^2)\log(4/\delta)}{n}} &= t
\end{aligned}$$

Then consider the difference in the loss.

$$\begin{aligned}
\ell(\theta_1) - \ell(\theta_2) &\le H|\theta_1 - \theta_0| \cdot |\theta_1 - \theta_2| && \text{(Lemma 1)} \\
&\le H \cdot \sqrt{\frac{\sigma_1^2 \log(4/\delta)}{n}} \cdot \sqrt{\frac{(\sigma_1^2 - \sigma_2^2)\log(4/\delta)}{n}} && \text{(with probability at least } 1 - \delta)
\end{aligned}$$

$$= \frac{H \log(4/\delta) \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}}{n} \qquad \text{(with probability at least } 1 - \delta)$$

Let $X = \ell(\theta_1) - \ell(\theta_2)$. Note that the predictive exposure is $\gamma = \mathbb{E}[X]$. Thus, using standard techniques to integrate the tail, we can convert the high-probability bound into a bound on the expectation, yielding the final bound on predictive exposure. We know

$$\Pr\left[X > \frac{H \log(4/\delta) \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}}{n}\right] \leq \delta.$$

Let $K = \frac{H \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}}{n}$, then

$$\Pr[X > t] \leq 4e^{-t/K}.$$

Let $t_0 = K \log(4)$. Thus,

$$\begin{aligned}
\mathbb{E}[X] &\leq \int_0^{t_0} 1dt + \int_{t_0}^{\infty} 4e^{-t/K} \\
&= \int_0^{K \log(4)} 1dt + \int_{K \log(4)}^{\infty} 4e^{-t/K} \\
&= K \log(4) + 4Ke^{-K \log(4)/K} \\
&= K \log(4) + 4Ke^{-\log(4)} \\
&= K \log(4) + K \\
&= K(1 + \log(4))
\end{aligned}$$

Thus,

$$\gamma = \frac{H(1 + \log(4)) \cdot \sigma_1 \sqrt{\sigma_1^2 - \sigma_2^2}}{n}.$$

In particular, we consider $n, N \to \infty$, $n/N \to 0$, so we do not need to consider a finite sample correction when estimating $E_{x \sim \mathrm{Unif}(S \setminus C(S))}[q(x)]$.

Now we compute the exact form of σ_1 and σ_2 using $[G^\top \Omega^{-1} G]^{-1}$. For $\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}$, we have $\theta = (p, \lambda)$ and

$$f_{C(S)}(x, \theta) = [w(x)(q(x) - p)].$$

We need to compute $[G^\top \Omega^{-1} G]^{-1}$.

$$\begin{aligned}
G &= \mathbb{E}\left[\tfrac{\partial f(x,\theta)}{\partial \theta}\right] \\
&= \mathbb{E}[-w(x)] \\
&= -1
\end{aligned}$$

$$\begin{aligned}
\Omega^{-1} &= \mathbb{E}\left[f(x,\theta) f(x,\theta)^\top\right]^{-1} \\
&= \mathbb{E}\left[\left[w(x)^2 (q(x) - p)^2\right]\right]^{-1} \\
&= \left[\mathrm{Var}[w(x)q(x)]\right]^{-1} \\
&= \frac{1}{\mathrm{Var}[w(x)q(x)]}
\end{aligned}$$

$$\begin{aligned}
[G(\theta_0)^\top \Omega^{-1}(\theta_0) G(\theta_0)]^{-1} &= \frac{1}{-1 \cdot \frac{1}{\mathrm{Var}[w(x)q(x)]} \cdot -1} \\
&= \mathrm{Var}[w(x)q(x)]
\end{aligned}$$

Then,

$$\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}] = \frac{\mathrm{Var}[w(x)q(x)]}{n}.$$

Thus, $\sigma_1^2 = \mathrm{Var}[w(x)q(x)]$.

Next, we establish the variance of optimal predictor $\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}$, where $\theta = (p, \lambda)$ and

$$f_{C(S),A(S)}(x,\theta) = \begin{bmatrix} w(x)(q(x)-p) \\ w(x)g(x,\lambda) \\ \tilde{\lambda}-\lambda \end{bmatrix}.$$

We need to compute $[G^\top \Omega^{-1} G]^{-1}$.

$$\begin{aligned} G &= \mathbb{E}\left[\tfrac{\partial f(x,\theta)}{\partial \theta}\right] \\ &= \mathbb{E}\begin{bmatrix} -w(x) & 0 \\ 0 & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ 0 & -I_d \end{bmatrix} \\ &= \mathbb{E}\begin{bmatrix} -1 & 0 \\ 0 & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ 0 & -I_d \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \Omega^{-1} &= \mathbb{E}\left[f(x,\theta)f(x,\theta)^\top\right]^{-1} \\ &= \mathbb{E}\left[\begin{bmatrix} w(x)^2(q(x)-p)^2 & w(x)^2(q(x)-p)g(x,\lambda)^\top & w(x)(q(x)-p)(\tilde{\lambda}-\lambda)^\top \\ w(x)^2 g(x,\lambda)(q(x)-p) & w(x)^2 g(x,\lambda)g(x,\lambda)^\top & w(x)g(x,\lambda)(\tilde{\lambda}-\lambda)^\top \\ w(x)(\tilde{\lambda}-\lambda)(q(x)-p) & w(x)(\tilde{\lambda}-\lambda)g(x,\lambda)^\top & (\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top \end{bmatrix}\right]^{-1} \\ &= \begin{bmatrix} \mathrm{Var}[w(x)q(x)] & \mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top] & 0 \\ \mathbb{E}[w(x)^2 g(x,\lambda)(q(x)-p)] & \mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top] \end{bmatrix}^{-1} \end{aligned}$$

Let $U^{-1} = \begin{bmatrix} \mathrm{Var}[w(x)q(x)] & \mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top] \\ \mathbb{E}[w(x)^2 g(x,\lambda)(q(x)-p)] & \mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top] \end{bmatrix}^{-1}$.

$$\Omega^{-1} = \begin{bmatrix} U^{-1}[1,1] & U^{-1}[1,2] & 0 \\ U^{-1}[2,1] & U^{-1}[2,2] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix}$$

Putting it together,

$$\begin{aligned} G^\top \Omega^{-1} G &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top & -I_d \end{bmatrix} \begin{bmatrix} U^{-1}[1,1] & U^{-1}[1,2] & 0 \\ U^{-1}[2,1] & U^{-1}[2,2] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ 0 & -I_d \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top & -I_d \end{bmatrix} \begin{bmatrix} -U^{-1}[1,1] & U^{-1}[1,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ -U^{-1}[2,1] & U^{-1}[2,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ 0 & -\mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \\ &= \begin{bmatrix} U^{-1}[1,1] & -U^{-1}[1,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ -\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top U^{-1}[2,1] & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top U^{-1}[2,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] + \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \\ &= \begin{bmatrix} U^{-1}[1,1] & -U^{-1}[1,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] \\ -\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top U^{-1}[2,1] & \mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}]^\top U^{-1}[2,2]\mathbb{E}[\frac{\partial w(x)g(x,\lambda)}{\partial \lambda}] + \mathrm{Var}[\Delta]^{-1} \end{bmatrix} \end{aligned}$$

Using Schur's complement, we can compute the entries of U^{-1}. Let $U = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, then

$U^{-1}[1,1] = (A - BD^{-1}C)^{-1}$

$$= \frac{1}{\mathrm{Var}[w(x)q(x)] - \mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]}$$

$$U^{-1}[1,2] = -(A - BD^{-1}C)^{-1}BD^{-1} = -U^{-1}[1,1]\cdot\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}$$

$$U^{-1}[2,1] = -D^{-1}C(A - BD^{-1}C)^{-1} = -U^{-1}[1,1]\cdot\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]$$

$$\begin{aligned}U^{-1}[2,2] &= D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}\\ &= \mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1} + (U^{-1}[1,1]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]\\ &\quad\cdot\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}).\end{aligned}$$

The $[1,1]$ entry of $[G^\top\Omega^{-1}G]^{-1}$ characterizes the variance of $\hat{p}$. For a matrix $M = \begin{bmatrix}A & B\\ C & D\end{bmatrix}$, $M^{-1}[1,1] = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$, using Schur's complement.

$$\begin{aligned}[G^\top\Omega^{-1}G]^{-1}[1,1] &= \frac{1}{U^{-1}[1,1]} + \frac{1}{U^{-1}[1,1]^2}U^{-1}[1,2]\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\Bigg(\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top U^{-1}[2,2]\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\\ &\qquad + \mathrm{Var}[\Delta]^{-1} - \frac{1}{U^{-1}[1,1]}\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top U^{-1}[2,1]U^{-1}[1,2]\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\Bigg)^{-1}\\ &\qquad\cdot\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top U^{-1}[2,1]\\ &= \frac{1}{U^{-1}[1,1]} + \frac{1}{U^{-1}[1,1]^2}\cdot U^{-1}[1,2]\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\\ &\qquad\cdot\Bigg(\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big] + \mathrm{Var}[\Delta]^{-1}\Bigg)^{-1}\\ &\qquad\cdot\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top U^{-1}[2,1]\\ &= \frac{1}{U^{-1}[1,1]} + \mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\\ &\qquad\cdot\Bigg(\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\Big[\tfrac{\partial g(x,\lambda)}{\partial\lambda}\Big] + \mathrm{Var}[\Delta]^{-1}\Bigg)^{-1}\\ &\qquad\cdot\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]\\ &= \frac{1}{U^{-1}[1,1]} + \mathrm{Var}[w(x)q(x)]\cdot c_0(w,q,g,\Delta)\\ &= \mathrm{Var}[w(x)q(x)]\\ &\qquad - \mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]\\ &\qquad + \mathrm{Var}[w(x)q(x)]\cdot c_0(w,q,g,\Delta)\end{aligned}$$

where

$$\begin{aligned}c_0(w,q,g,\Delta) &= \frac{1}{\mathrm{Var}[w(x)q(x)]}\cdot\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]\\ &\qquad\cdot\Bigg(\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big] + \mathrm{Var}[\Delta]^{-1}\Bigg)^{-1}\\ &\qquad\cdot\mathbb{E}\Big[\tfrac{\partial w(x)g(x,\lambda)}{\partial\lambda}\Big]^\top\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)].\end{aligned}$$

Recall from Definition 4,

$$\begin{aligned}\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda)) &= \max_{c\in\mathbb{R}^d}\rho(w(x)q(x), c^\top w(x)g(x,\lambda))\\ &= \max_{c\in\mathbb{R}^d}\frac{\mathrm{Cov}(w(x)q(x), c^\top w(x)g(x,\lambda))}{\sqrt{\mathrm{Var}[w(x)q(x)]\cdot\mathrm{Var}[c^\top w(x)g(x,\lambda)]}}\end{aligned}$$

$$= \max_{c \in \mathbb{R}^d} \frac{c^\top \mathrm{Cov}(w(x)q(x), w(x)g(x,\lambda))}{\sqrt{\mathrm{Var}[w(x)q(x)] \cdot c^\top \mathrm{Cov}(w(x)g(x,\lambda), w(x)g(x,\lambda))c}}.$$

Squaring the objective,

$$\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 = \max_{c \in \mathbb{R}^d} \frac{c^\top \mathrm{Cov}(w(x)g(x,\lambda), w(x)q(x))\mathrm{Cov}(w(x)g(x,\lambda), w(x)q(x))^\top c}{\mathrm{Var}[q(x)] \cdot c^\top \mathrm{Cov}(w(x)g(x,\lambda), w(x)g(x,\lambda))c}$$

This is a generalized Rayleigh quotient with solution

$$\begin{aligned}
&\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 \\
&\quad = \frac{\mathrm{Cov}(w(x)g(x,\lambda), w(x)q(x))^\top \mathrm{Cov}(w(x)g(x,\lambda), w(x)g(x,\lambda))^{-1} \mathrm{Cov}(w(x)g(x,\lambda), w(x)q(x))}{\mathrm{Var}[w(x)q(x)]} \\
&\quad = \frac{\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)^\top]\mathbb{E}[w(x)^2 g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[w(x)^2(q(x)-p)g(x,\lambda)]}{\mathrm{Var}[w(x)q(x)]}.
\end{aligned}$$

Thus,

$$\begin{aligned}
[G^\top \Omega^{-1} G]^{-1}[1,1] &= \mathrm{Var}[w(x)q(x)] - \mathrm{Var}[w(x)q(x)] \cdot \rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 + \mathrm{Var}[w(x)q(x)] \cdot c_0(w,q,g,\Delta) \\
&= \mathrm{Var}[w(x)q(x)] \cdot (1 - \rho_{\mathrm{cc}}(q(x), g(x,\lambda))^2 + c_0(w,q,g,\Delta))
\end{aligned}$$

Then,

$$\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}] = \frac{\mathrm{Var}[w(x)q(x)](1 - \rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 + c_0(w,q,g,\Delta))}{n}.$$

Thus, $\sigma_2^2 = \mathrm{Var}[w(x)q(x)](1 - \rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 + c_0(w,q,g,\Delta))$.

$$\begin{aligned}
\sigma_1\sqrt{\sigma_1^2 - \sigma_2^2} &= \sqrt{\mathrm{Var}[w(x)q(x)]} \cdot \sqrt{\mathrm{Var}[w(x)q(x)] - \mathrm{Var}[w(x)q(x)](1 - \rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(w,q,g,\Delta))} \\
&= \sqrt{\mathrm{Var}[w(x)q(x)]} \cdot \sqrt{\mathrm{Var}[w(x)q(x)](\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(w,q,g,\Delta))} \\
&= \mathrm{Var}[w(x)q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(w,q,g,\Delta)}
\end{aligned}$$

Plugging the above into the predictive exposure bound,

$$\gamma \leq \frac{H(1+\log(4)) \cdot \mathrm{Var}[w(x)q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(w,q,g,\Delta)}}{n}$$

where each term is evaluated at the true p_0 and λ_0 □

Theorem 12 follows analogously by plugging in $w'(x)(q(x)-p)$ in place of $w(x)(q(x)-p)$ is the proof above.

Theorem 22. *Let S be a dataset of size N and let $C(S)$ be random sample without replacement of size $n = \alpha N$, $0 < \alpha < 1$. For any dataset S, given moment conditions $f_{K(S)}(x,\theta)$, the optimal predictor $\hat{\theta}^{\mathrm{GMM}}_{q|K(S)}$ of $p_{S \setminus C(S)} = \mathbb{E}_{x \sim \mathrm{Unif}(S \setminus C(S))}[q(x)]$ has asymptotic distribution*

$$\sqrt{(1-\alpha)\alpha N}(\hat{\theta}^{\mathrm{GMM}}_{q|K(S)} - \theta_0) \sim \mathcal{N}(0, [G(\theta_0)^\top \Omega^{-1}(\theta_0) G(\theta_0)]^{-1}).$$

Proof. The optimal predictor $\hat{\theta}^{\mathrm{GMM}}$ of $p = \mathbb{E}_{\mathrm{Unif}(S)}[q(x)]$ is

$$\sqrt{\frac{\alpha N}{1-\alpha}}(\hat{\theta}^{\mathrm{GMM}} - \theta_0) \sim \mathcal{N}(0, [G(\theta_0)^\top \Omega^{-1}(\theta_0) G(\theta_0)]^{-1}),$$

by plugging in $n = \alpha N$ and n appropriately into Theorem 20.

For any fixed sample $C(S)$, an estimate for p implies an estimate for $p_{S \setminus C(S)}$ using a simple translation and scaling, $p = (1-\alpha)p_{S \setminus C(S)} + \alpha \mathbb{E}_{x \sim \mathrm{Unif}(C(S))}[q(x)]$. $\mathrm{Var}[\hat{p}_{S \setminus C(S)}] = \frac{1}{(1-\alpha)^2} \cdot \mathrm{Var}[\hat{p}]$. Thus, the optimal predictor $\hat{\theta}^{\mathrm{GMM}}_{q|K(S)}$ of $p_{S \setminus C(S)} = \mathbb{E}_{\mathrm{Unif}(S \setminus C(S))}[q(x)]$ is

$$\sqrt{(1-\alpha)\alpha N}(\hat{\theta}^{\mathrm{GMM}}_{q|K(S)} - \theta_0) \sim \mathcal{N}(0, [G(\theta_0)^\top \Omega^{-1}(\theta_0) G(\theta_0)]^{-1}).$$

□

Corollary 1. *Let $A(S) = \tilde{\lambda} = \lambda + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. As $n, N \to \infty$, $n/N \to \alpha$, for a convex, H-smooth ℓ, $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) \le \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(x), g(x,\lambda))^2 - c_0(1, q, g, \Delta)}$.*

Proof. We would like to find the efficient GMM estimator for $p_{S \setminus C(S)} = \mathbb{E}_{x \sim \mathrm{Unif}(S \setminus C(S))}[q(x_i)]$. Typical GMM for sampling without replacement finds the efficient estimator for $p_S = \mathbb{E}_{x \sim \mathrm{Unif}(S)}[q(x_i)]$. We find that the efficient GMM estimator for p_S implies an efficient GMM estimator for $p_{S \setminus C(S)}$. We use Theorem 22 to find the efficient estimator $\hat{\theta}^{\mathrm{GMM}}_{q|K(S)}$ for $p_{S \setminus C(S)}$. Theorem 22 essentially performs a finite sample correction for estimation.

From Theorem 22, we have

$$\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S)}] = \frac{\sigma_1}{(1-\alpha)\alpha N}$$

and

$$\mathrm{Var}[\hat{\theta}^{\mathrm{GMM}}_{q|C(S),A(S)}] = \frac{\sigma_2}{(1-\alpha)\alpha N}.$$

Following the proof of Theorem 4, we have $\sigma_1^2 = \mathrm{Var}[w(x)q(x)]$, $\sigma_2^2 = \mathrm{Var}[w(x)q(x)](1 - \rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 + c_0(w, q, g, \Delta))$, and

$$\sigma_1\sqrt{\sigma_1^2 - \sigma_2^2} = \mathrm{Var}[w(x)q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(w, q, g, \Delta)}$$

We plug in $w(x) = 1$ to get

$$\sigma_1\sqrt{\sigma_1^2 - \sigma_2^2} = \mathrm{Var}[q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(x), g(x,\lambda))^2 - c_0(1, q, g, \Delta)}$$

Thus, we have

$$\gamma \le \frac{H(1+\log(4)) \cdot \mathrm{Var}[q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(x), g(x,\lambda))^2 - c_0(1, q, g, \Delta)}}{(1-\alpha)\alpha N}$$

□

A.10 Proofs from Section 5

A.10.1 Theorem 5, Dataset Statistics

Theorem 5. *Let $A(S) = \tilde{\lambda} = \frac{1}{N}\sum_{i=1}^N \Gamma(x_i) + \Delta$, $\Delta \sim \mathcal{N}(0, \sigma_\Delta)$. Let $g(x,\lambda) = \Gamma(x) - \lambda$. For a convex, H-smooth ℓ, $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) = \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(x)] \cdot \rho(q(x), \Gamma(x)) \cdot \sqrt{\frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)] + \sigma_\Delta}}$.*

Proof. From Corollary 1, we know

$$\begin{aligned} c_0(1, q, g, \Delta) = \frac{1}{\mathrm{Var}[q(x)]} &\cdot \mathbb{E}[(q(x) - p)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\tfrac{\partial g(x,\lambda)}{\partial \lambda}\right] \\ &\cdot \left(\mathbb{E}\left[\tfrac{\partial g(x,\lambda)}{\partial \lambda}\right]^\top \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\tfrac{\partial g(x,\lambda)}{\partial \lambda}\right] + \mathrm{Var}[\Delta]^{-1}\right)^{-1} \\ &\cdot \mathbb{E}\left[\tfrac{\partial g(x,\lambda)}{\partial \lambda}\right]^\top \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[(q(x) - p)g(x,\lambda)]. \end{aligned}$$

When $g(x,\lambda) = \Gamma(x) - \lambda$, $\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] = 1$. Also notice that $\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top] = \mathbb{E}[(\Gamma(x) - \lambda)(\Gamma(x) - \lambda)^\top] = \mathrm{Var}[\Gamma(x)]$. Thus,

$$\begin{aligned} c_0(1, q, g, \Delta) = \frac{1}{\mathrm{Var}[q(x)]} &\cdot \mathbb{E}[(q(x) - p)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\left(\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1} + \mathrm{Var}[\Delta]^{-1}\right)^{-1} \\ &\cdot \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[(q(x) - p)g(x,\lambda)] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mathrm{Var}[q(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)^\top]\mathrm{Var}[\Gamma(x)]^{-1}\left(\mathrm{Var}[\Gamma(x)]^{-1}+\mathrm{Var}[\Delta]^{-1}\right)^{-1}\mathrm{Var}[\Gamma(x)]^{-1} \\
&\quad \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)] \\
&= \frac{1}{\mathrm{Var}[q(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)^\top] \cdot \frac{1}{\mathrm{Var}[\Gamma(x)]}\left(\frac{1}{\mathrm{Var}[\Gamma(x)]}+\frac{1}{\mathrm{Var}[\Delta]}\right)^{-1}\frac{1}{\mathrm{Var}[\Gamma(x)]} \\
&\quad \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)] \\
&= \frac{1}{\mathrm{Var}[q(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)^\top] \cdot \frac{1}{\mathrm{Var}[\Gamma(x)]}\left(\frac{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]\mathrm{Var}[\Delta]}\right)^{-1}\frac{1}{\mathrm{Var}[\Gamma(x)]} \\
&\quad \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)] \\
&= \frac{1}{\mathrm{Var}[q(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)^\top] \cdot \frac{1}{\mathrm{Var}[\Gamma(x)]}\left(\frac{\mathrm{Var}[\Gamma(x)]\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}\right)\frac{1}{\mathrm{Var}[\Gamma(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)] \\
&= \frac{1}{\mathrm{Var}[q(x)]} \cdot \mathbb{E}[(q(x)-p)g(x,\lambda)^\top] \cdot \frac{1}{\mathrm{Var}[\Gamma(x)]}\left(\frac{\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}\right)\mathbb{E}[(q(x)-p)g(x,\lambda)] \\
&= \frac{\mathrm{Cov}(q(x)-p, g(x,\lambda))}{\mathrm{Var}[q(x)]\mathrm{Var}[\Gamma(x)]} \cdot \frac{\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]} \\
&= \frac{\mathrm{Cov}(q(x), \Gamma(x))}{\mathrm{Var}[q(x)]\mathrm{Var}[\Gamma(x)]} \cdot \frac{\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]} \\
&= \rho(q(x),\Gamma(x))^2 \cdot \frac{\mathrm{Var}[\Delta]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]} \\
&= \rho(q(x),\Gamma(x))^2 \cdot \left(1-\frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}\right)
\end{aligned}
$$

Plugging the above into Corollary 1 along with $\rho_{\mathrm{cc}}(q(x), g(x,\lambda))^2 = \rho(q(x),\Gamma(x))^2$,

$$
\begin{aligned}
\gamma &= \frac{H(1+\log(4)) \cdot \mathrm{Var}[q(x)] \cdot \sqrt{\rho(q(x),\Gamma(x))^2 - \rho(q(x),\Gamma(x))^2 \cdot \left(1-\frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}\right)}}{(1-\alpha)\alpha N} \\
&= \frac{H(1+\log(4)) \cdot \mathrm{Var}[q(x)] \cdot \sqrt{\rho(q(x),\Gamma(x))^2 \cdot \frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}}}{(1-\alpha)\alpha N} \\
&= \frac{H(1+\log(4)) \cdot \mathrm{Var}[q(x)] \cdot \rho(q(x),\Gamma(x)) \cdot \sqrt{\frac{\mathrm{Var}[\Gamma(x)]}{\mathrm{Var}[\Gamma(x)]+\mathrm{Var}[\Delta]}}}{(1-\alpha)\alpha N}
\end{aligned}
$$

□

A.10.2 Theorem 6, Empirical Risk Minimization

Theorem 6. *Let* $w_{\mathrm{ERM}} = \arg\min_w \frac{1}{N}\sum_{i=1}^N \ell(z_i, w)$*, for* ℓ *strictly convex and* $\mathbb{E}[\nabla^2 \ell(z, w_{\mathrm{ERM}})]$ *positive definite. Let* $A(S) = w_{\mathrm{ERM}} + \Delta$*,* $\Delta \sim \mathcal{N}(0, \Sigma_\Delta)$*. Let* $g(z, w_{\mathrm{ERM}}) = \nabla \ell(z, w_{\mathrm{ERM}})$*. For a convex,* H*-smooth* ℓ*,* $\gamma(\mathcal{P}_{\mathrm{RS}}, q, \ell, A) \leq \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q(z)] \cdot \sqrt{\rho_{\mathrm{cc}}(q(z), g(z, w_{\mathrm{ERM}}))^2 - c_0(w, q, g, \Delta)}$.

Proof. The result directly follows by plugging in $g(z, w_{\mathrm{ERM}}) = \nabla \ell(z, w_{\mathrm{ERM}}) = \nabla_z$ into Corollary 1 with $\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] = \nabla^2 \ell(x, w_{\mathrm{ERM}}) = H_z$.

In particular, $c_0(1, q, g, \Delta) = \frac{1}{\mathrm{Var}[q(z)]} \cdot \mathbb{E}[(q(x)-p)\nabla_z^\top]\mathbb{E}[\nabla_z \nabla_z^\top]^{-1}\mathbb{E}[H_z](\mathbb{E}[H_z]\mathbb{E}[\nabla_z \nabla_z^\top]^{-1}$ $\mathbb{E}[H_z] + \mathrm{Var}[\Delta]^{-1})^{-1}\mathbb{E}[H_z]\mathbb{E}[\nabla_z \nabla_z^\top]^{-1}\mathbb{E}[(q(x)-p)\nabla_z]$

□

A.10.3 Theorem 7, Calibrated Noise Scheme

Theorem 7. *Let* $\nabla_z = \nabla \ell(z, w_{\mathrm{ERM}})$ *and* $H_z = \nabla^2 \ell(z, w_{\mathrm{ERM}})$. *Let* $A(S) = \tilde{w} = w_{\mathrm{ERM}} + \Delta$,

$$\Delta \sim \mathcal{N}(0, \sigma^2 \mathbb{E}[w(x) H_z]^{-1} \mathbb{E}[w(x)^2 \nabla_z \nabla_z^\top] \mathbb{E}[w(x) H_z]^{-1}).$$

For a convex, H-smooth ℓ, $\gamma(\mathcal{P}, q, \ell, A) \le \frac{H(1+\log(4))}{(1-\alpha)\alpha N} \cdot \mathrm{Var}[q'] \cdot \rho_{\mathrm{cc}}(q', g') \cdot \sqrt{\frac{1}{\sigma^2+1}}$.

Proof. Let $A = \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z]$. Then

$$c_0(w, q, g, \Delta) = \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z] \left(A + \mathbb{E}[(\tilde{w} - w_{\mathrm{ERM}})(\tilde{w} - w_{\mathrm{ERM}})^\top]^{-1} \right)^{-1} \cdot \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z]$$

and

$$\mathbb{E}[(\tilde{w} - w_{\mathrm{ERM}})(\tilde{w} - w_{\mathrm{ERM}})^\top]^{-1} = \frac{1}{\sigma^2} \cdot A$$

Plugging in $\mathbb{E}[(\tilde{w} - w_{\mathrm{ERM}})(\tilde{w} - w_{\mathrm{ERM}})^\top]^{-1}$, we have

$$\begin{aligned}
c_0(1, q, g, \Delta) &= \frac{1}{\mathrm{Var}[w(z) q(z)]} \cdot \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z] \left(A + \frac{1}{\sigma^2} \cdot A \right)^{-1} \\
&\quad \cdot \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z] \\
&= \frac{1}{\mathrm{Var}[w(z) q(z)]} \cdot \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z] \left(\left(1 + \frac{1}{\sigma^2} \right) A \right)^{-1} \\
&\quad \cdot \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z] \\
&= \frac{1}{\mathrm{Var}[w(z) q(z)]} \cdot \frac{1}{1 + \frac{1}{\sigma^2}} \cdot \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z] A^{-1} \\
&\quad \cdot \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z] \\
&= \frac{1}{\mathrm{Var}[w(z) q(z)]} \cdot \frac{\sigma^2}{\sigma^2 + 1} \cdot \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z) H_z] \mathbb{E}[w(z) H_z]^{-1} \\
&\quad \cdot \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top] [\mathbb{E}[w(z) H_z]^\top]^{-1} \mathbb{E}[w(z) H_z]^\top \mathbb{E}[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z] \\
&= \frac{1}{\mathrm{Var}[w(z) q(z)]} \cdot \frac{\sigma^2}{\sigma^2 + 1} \cdot \mathbb{E}[w(z)^2 (q(z) - p) \nabla_z^\top] \mathbb{E}[[w(z)^2 \nabla_z \nabla_z^\top]^{-1} \mathbb{E}[[w(z)^2 (q(z) - p) \nabla_z] \\
&= \frac{\sigma^2}{\sigma^2 + 1} \cdot \rho_{\mathrm{cc}}(q', g')^2
\end{aligned}$$

Plugging $c_0(1, q, g, \Delta)$ into the predictive exposure bound from Theorem 4,

$$\begin{aligned}
\gamma &= \frac{H(1 + \log(4)) \cdot \mathrm{Var}[q']}{(1 - \alpha)\alpha N} \cdot \sqrt{(\rho_{\mathrm{cc}}(q', g')^2 - c_0(1, q, g, \Delta))} \\
&= \frac{H(1 + \log(4)) \cdot \mathrm{Var}[q']}{(1 - \alpha)\alpha N} \cdot \sqrt{\rho_{\mathrm{cc}}(q', g')^2 - \frac{\sigma^2}{\sigma^2 + 1} \cdot \rho_{\mathrm{cc}}(q', g')^2} \\
&= \frac{H(1 + \log(4)) \cdot \mathrm{Var}[q']}{(1 - \alpha)\alpha N} \cdot \sqrt{\frac{\rho_{\mathrm{cc}}(q', g')^2}{\sigma^2 + 1}}.
\end{aligned}$$

□

A.10.4 Calibrated Noise and Linear Regression

Theorem 8. *Let* $\tilde{w} = w_{\mathrm{ERM}} + \Delta$, $\Delta \sim \mathcal{N}(0, \sigma^2 \cdot \mathbb{E}[xx^\top]^{-1} \mathbb{E}[(w_{\mathrm{ERM}}^\top x - y)^2 xx^\top] \mathbb{E}[xx^\top]^{-1})$. *Then*

$$\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} (\tilde{w}^\top x_i - w_{\mathrm{ERM}}^\top x_i)^2 \right] \le \min\left\{ \frac{\sigma^2 (\max_i \|x_i\|^2)}{\lambda_{\min}(\mathbb{E}[xx^\top])} \cdot \hat{\mathcal{L}}_S(w_{\mathrm{ERM}}), \sigma^2 d \cdot \hat{\mathcal{L}}_S^{\max}(w_{\mathrm{ERM}}) \right\}.$$

where $\hat{\mathcal{L}}_S(w_{\mathrm{ERM}}) = \frac{1}{N} \sum_{i=1}^{N} (w_{\mathrm{ERM}}^\top x_i - y_i)^2$ *and* $\hat{\mathcal{L}}_S^{\max}(w_{\mathrm{ERM}}) = \max_i (w_{\mathrm{ERM}}^\top x_i - y_i)^2$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\Delta^\top x_i)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Delta^\top x_i x_i^\top \Delta\right] \\
&= \mathbb{E}\left[\Delta^\top \mathbb{E}[xx^\top]\Delta\right] \\
&= \mathbb{E}\left[\operatorname{tr}(\Delta^\top \mathbb{E}[xx^\top]\Delta)\right] \\
&= \mathbb{E}\left[\operatorname{tr}(\mathbb{E}[xx^\top]\Delta\Delta^\top)\right] \\
&= \operatorname{tr}(\mathbb{E}[xx^\top]\cdot\mathbb{E}[\Delta\Delta^\top]) \\
&= \operatorname{tr}(\mathbb{E}[xx^\top]\cdot\sigma^2\mathbb{E}[xx^\top]^{-1}\mathbb{E}[(w_{\text{ERM}}^\top x - y)^2 xx^\top]\mathbb{E}[xx^\top]^{-1}) \\
&= \sigma^2\cdot\operatorname{tr}(\mathbb{E}[(w_{\text{ERM}}^\top x - y)^2 xx^\top]\mathbb{E}[xx^\top]^{-1}) \\
&= \sigma^2\cdot\mathbb{E}\left[\operatorname{tr}((w_{\text{ERM}}^\top x - y)^2 xx^\top\mathbb{E}[xx^\top]^{-1})\right] \\
&= \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot\operatorname{tr}(xx^\top\mathbb{E}[xx^\top]^{-1})\right] \\
&= \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot\operatorname{tr}(x^\top\mathbb{E}[xx^\top]^{-1}x)\right] \\
&= \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot x^\top\mathbb{E}[xx^\top]^{-1}x\right]
\end{aligned}$$

From here, we establish two upper bounds. First, we establish an upper bound based on the average empirical loss.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] &= \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot x^\top\mathbb{E}[xx^\top]^{-1}x\right] \\
&\le \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot\frac{\|x\|^2}{\lambda_{\min}(\mathbb{E}[xx^\top])}\right] \\
&\le \sigma^2\cdot\mathbb{E}\left[(w_{\text{ERM}}^\top x - y)^2\cdot\frac{\max_i\|x_i\|^2}{\lambda_{\min}(\mathbb{E}[xx^\top])}\right] \\
&\le \frac{\sigma^2\cdot(\max_i\|x_i\|^2)}{\lambda_{\min}(\mathbb{E}[xx^\top])}\cdot\mathbb{E}[(w_{\text{ERM}}^\top x - y)^2] \\
&= \frac{\sigma^2\cdot(\max_i\|x_i\|^2)}{\lambda_{\min}(\mathbb{E}[xx^\top])}\cdot\left(\frac{1}{N}\sum_{i=1}^{N}(w_{\text{ERM}}^\top x_i - y_i)^2\right)
\end{aligned}$$

Next, we establish an upper bound based on the maximum empirical loss.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - \hat{w}^\top x_i)^2\right] &= \sigma^2\cdot\mathbb{E}\left[(\hat{w}^\top x - y)^2\cdot x^\top\mathbb{E}[xx^\top]^{-1}x\right] \\
&\le \left(\max_i(\hat{w}^\top x_i - y_i)^2\right)\cdot\sigma^2\cdot\mathbb{E}\left[x^\top\mathbb{E}[xx^\top]^{-1}x\right] \\
&= \left(\max_i(\hat{w}^\top x_i - y_i)^2\right)\cdot\sigma^2\cdot\mathbb{E}\left[\operatorname{tr}(\mathbb{E}[xx^\top]^{-1}xx^\top)\right] \\
&= \left(\max_i(\hat{w}^\top x_i - y_i)^2\right)\cdot\sigma^2\cdot\operatorname{tr}(\mathbb{E}[xx^\top]^{-1}\mathbb{E}[xx^\top]) \\
&= \left(\max_i(\hat{w}^\top x_i - y_i)^2\right)\cdot\sigma^2\cdot\operatorname{tr}(I_d) \\
&= \left(\max_i(\hat{w}^\top x_i - y_i)^2\right)\cdot\sigma^2\cdot d
\end{aligned}$$

□

Lemma 2. *When* $\Delta\sim\mathcal{N}(0,\sigma^2 I_d)$, $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N}\|x_i\|^2\right)$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{w}^\top x_i - w_{\text{ERM}}^\top x_i)^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\Delta^\top x_i)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Delta^\top x_i x_i^\top \Delta\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(\Delta^\top x_i x_i^\top \Delta)\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(\Delta\Delta^\top x_i x_i^\top)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(\mathbb{E}[\Delta\Delta^\top]\cdot x_i x_i^\top) \\
&= \frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(\sigma^2 I_d \cdot x_i x_i^\top) \\
&= \sigma^2\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(x_i x_i^\top)\right) \\
&= \sigma^2\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}(x_i^\top x_i)\right) \\
&= \sigma^2\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\|x_i\|^2\right)
\end{aligned}$$

□

A.11 Proofs from Section 6

A.11.1 Theorem 9, Finite Family of Queries and Processes

Theorem 9. *Let $A(S) = \tilde{\lambda} = \lambda_0 + \Delta$, $\Delta = \mathcal{N}(0, \Sigma_\Delta)$. For a finite family of queries $\mathcal{Q}$, a finite family of processes $\boldsymbol{\mathcal{P}}$ satisfying Assumption 1, and a convex, H-smooth ℓ, the predictive exposure is* $\sup_{\mathcal{P}\in\boldsymbol{\mathcal{P}}}\gamma(\mathcal{P},\mathcal{Q},\ell,A) \le \frac{H(1+\log(4|\mathcal{Q}||\mathcal{W}_{\boldsymbol{\mathcal{P}}}|))}{n}\ \sup_{w\in\mathcal{W}_{\boldsymbol{\mathcal{P}}}}\sup_{q\in\mathcal{Q}}\operatorname{Var}[q']\sqrt{\rho_{\text{cc}}(q',g')^2 - c_0(w,q,g,\Delta)}$

Proof. Let $n = (1-\alpha)\alpha N$. For a weight $w \in \mathcal{W}_{\boldsymbol{\mathcal{P}}}$ and query $q \in \mathcal{Q}$, let $X_{w,q} = \ell(\hat{\theta}^{\text{GMM}}_{q|C(S)}) - \ell(\hat{\theta}^{\text{GMM}}_{q|C(S),A(S)})$. From Theorem 4, we know

$$\Pr\left[X_{w,q} > \frac{H\log(4/\delta)\cdot\sigma_{w,q,1}\sqrt{\sigma^2_{w,q,1}-\sigma^2_{w,q,2}}}{n}\right] \le \delta.$$

Furthermore, every $X_{w,q}$ satisfies

$$\Pr\left[X_{w,q} > \frac{H\log(4/\delta)\cdot\sup_{q\in\mathcal{Q}}\sup_{w\in\mathcal{W}_{\boldsymbol{\mathcal{P}}}}\sigma_{w,q,1}\sqrt{\sigma^2_{w,q,1}-\sigma^2_{w,q,2}}}{n}\right] \le \delta.$$

Let $Y = \sup_{q\in\mathcal{Q}}\sup_{w\in\mathcal{W}_{\boldsymbol{\mathcal{P}}}} X_{w,q}$. Then, using a union bound, with probability at least $1-\delta$,

$$\Pr\left[Y > \frac{H\log(4|\mathcal{Q}||\mathcal{W}_{\boldsymbol{\mathcal{P}}}|/\delta)\cdot\sup_{q\in\mathcal{Q}}\sup_{w\in\mathcal{W}_{\boldsymbol{\mathcal{P}}}}\sigma_{w,q,1}\sqrt{\sigma^2_{w,q,1}-\sigma^2_{w,q,2}}}{n}\right] \le \delta.$$

Now, consider $\mathbb{E}[Y]$. Thus, using standard techniques to integrate the tail, we can convert the high-probability bound into a bound on the expectation, yielding the final predictive exposure bound.

Following the proof from Theorem 4, we can conclude that

$$\mathbb{E}[Y] \le \frac{H(1+\log(4|\mathcal{Q}||\mathcal{W}_{\boldsymbol{\mathcal{P}}}|)) \cdot \sup_{q\in\mathcal{Q}} \sigma_{w,q,1}\sqrt{\sigma_{w,q,1}^2 - \sigma_{w,q,2}^2}}{n}$$

where $n = (1-\alpha)\alpha N$.

From Theorem 4, we have that

$$\sigma_{w,q,1}^2 = \mathrm{Var}[w(x)q(x)],$$
$$\sigma_{w,q,1}^2 - \sigma_{w,q,2}^2 = \mathrm{Var}[w(x)q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(1,q,g,\Delta)}$$

Thus, we can conclude that

$$\mathbb{E}[Y] \le \frac{H(1+\log(4|\mathcal{Q}||\mathcal{W}_{\boldsymbol{\mathcal{P}}}|)) \cdot \sup_{q\in\mathcal{Q}} \mathrm{Var}[q(x)] \cdot \sqrt{\rho_{\mathrm{cc}}(w(x)q(x), w(x)g(x,\lambda))^2 - c_0(1,q,g,\Delta)}}{n}.$$

Lastly, note that

$$\sup_{\mathcal{P}\in\boldsymbol{\mathcal{P}}} \gamma(\mathcal{P},\mathcal{Q},\ell,A) \le \mathbb{E}[Y]$$

by pulling out the supremum using Jensen's inequality. □

A.11.2 Theorem 10, Linear Queries

Definition 14 (Gaussian Complexity (Bartlett & Mendelson, 2003))**.** *Let $\xi \sim \mathcal{N}(0, I_d)$. The Gaussian complexity of a set K is defined as $\mathcal{G}(K) = \mathbb{E}\left[\sup_{u\in K}\langle u, \xi\rangle\right]$.*

Theorem 23. *Let query family $\mathcal{Q}_{\mathrm{lin}} = \{q(x) = u^\top x \mid u \in \mathcal{U}_{\mathrm{lin}} \subseteq \mathbb{R}^d\}$. Let diameter $\mathcal{D}_\Sigma(K) = \sup_{u\in K} u^\top \Sigma u$. Let $\mathcal{U}_{\mathrm{lin},\Sigma} = \{\Sigma^{1/2}u \mid u \in \mathcal{U}_{\mathrm{lin}}\}$. Using moment conditions*

$$f_{C(S)}(x,\theta) = [x-\mu] \text{ and } f_{C(S),A(S)}(x,\theta) = [x-\mu, g(x,\lambda), \tilde{\lambda}-\lambda]^\top, \tag{2}$$

we have

$$\sqrt{(1-\alpha)\alpha N}(\hat{\theta}^{\mathrm{GMM}}_{x|C(S)} - \mu) \sim \mathcal{N}(0,\Sigma_1) \text{ and } \sqrt{(1-\alpha)\alpha N}(\hat{\theta}^{\mathrm{GMM}}_{x|C(S),A(S)} - \mu) \sim \mathcal{N}(0,\Sigma_2)$$

where

$$\Sigma_1 = \mathrm{Var}[x],$$
$$\Sigma_2 = \mathrm{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] - c_0(1,x,g,\Delta).$$

and $c_0(1,x,g,\Delta) = \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\left(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\right.$
$\left.\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] + \Sigma_\Delta^{-1}\right)^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top].$

The predictive exposure of an algorithm A with respect to $\mathcal{Q}_{\mathrm{lin}}$ under a convex, H-smooth loss ℓ is

$$\gamma \le \frac{H}{(1-\alpha)\alpha N}\left(\mathcal{G}(\mathcal{U}_{\mathrm{lin},\Sigma_1}) + \sqrt{2(1+\log(2))\cdot\mathcal{D}_{\Sigma_1}(\mathcal{U}_{\mathrm{lin}})}\right)\left(\mathcal{G}(\mathcal{U}_{\mathrm{lin},\Sigma_1-\Sigma_2}) + \sqrt{2(1+\log(2))\cdot\mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\mathrm{lin}})}\right).$$

Proof. For each $u \in \mathcal{U}_{\mathrm{lin}}$, $\hat{\theta}^{\mathrm{GMM}}_{u|C(S)}$ estimates $\mathbb{E}[u^\top x \mid C(S)]$ and $\hat{\theta}^{\mathrm{GMM}}_{u|C(S),A(S)}$ estimates $\mathbb{E}[u^\top x \mid C(S), A(S)]$. Consider $\hat{\theta}^{\mathrm{GMM}}_{x|C(S)}$ which estimates $\mathbb{E}[x \mid C(S)]$ and $\hat{\theta}^{\mathrm{GMM}}_{x|C(S),A(S)}$ estimates $\mathbb{E}[x \mid C(S), A(S)]$.

We want to bound $\sup_{u\in\mathcal{U}_{\mathrm{lin}}} \ell(\hat{\theta}^{\mathrm{GMM}}_{u|C(S)}) - \ell(\hat{\theta}^{\mathrm{GMM}}_{u|C(S),A(S)})$ where ℓ is the expected log loss.

$$\sup_{u\in\mathcal{U}_{\mathrm{lin}}} \ell(\hat{\theta}^{\mathrm{GMM}}_{u|C(S)}) - \ell(\hat{\theta}^{\mathrm{GMM}}_{u|C(S),A(S)}) \le \sup_{u\in\mathcal{U}_{\mathrm{lin}}} H\cdot|\hat{\theta}^{\mathrm{GMM}}_{u|C(S)} - \theta_0^u|\cdot|\hat{\theta}^{\mathrm{GMM}}_{u|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{u|C(S),A(S)}|$$
$$= \sup_{u\in\mathcal{U}_{\mathrm{lin}}} H\cdot|u^\top(\hat{\theta}^{\mathrm{GMM}}_{x|C(S)} - \mu)|\cdot|u^\top(\hat{\theta}^{\mathrm{GMM}}_{x|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{x|C(S),A(S)})|$$

From GMM and Theorem 3, we know that $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu) \sim \mathcal{N}(0, \Sigma_1)$, $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)} - \mu) \sim \mathcal{N}(0, \Sigma_2)$, and $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \hat{\theta}^{\text{GMM}}_{x|C(S),A(S)}) \sim \mathcal{N}(0, \Sigma_1 - \Sigma_2)$, $n = (1-\alpha)\alpha N$. Using Theorem 25, we can bound the $|u^\top(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu)|$ and $|u^\top(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \hat{\theta}^{\text{GMM}}_{x|C(S),A(S)})|$ with high probability. Thus, with probability at least $1-\delta$,

$$\begin{aligned}
\sup_{u\in\mathcal{U}_{\text{lin}}} \ell(\hat{\theta}^{\text{GMM}}_{u|C(S)}) - \ell(\hat{\theta}^{\text{GMM}}_{u|C(S),A(S)}) &\le \sup_{u\in\mathcal{U}_{\text{lin}}} H \cdot |u^\top(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu)| \cdot |u^\top(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \hat{\theta}^{\text{GMM}}_{x|C(S),A(S)})| \\
&\le \frac{H}{n} \cdot \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1}) + \sqrt{2\log(2/\delta) \cdot \sup_{u\in\mathcal{U}_{\text{lin}}} u^\top \Sigma_1 u}\right) \\
&\quad \cdot \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1-\Sigma_2}) + \sqrt{2\log(2/\delta) \cdot \sup_{u\in\mathcal{U}_{\text{lin}}} u^\top (\Sigma_1 - \Sigma_2) u}\right) \\
&= \frac{H}{n} \cdot \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1}) + \sqrt{2\log(2/\delta) \cdot \mathcal{D}_{\Sigma_1}(\mathcal{U}_{\text{lin}})}\right) \\
&\quad \cdot \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1-\Sigma_2}) + \sqrt{2\log(2/\delta) \cdot \mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\text{lin}})}\right).
\end{aligned}$$

Let $X = \sup_{u\in\mathcal{U}_{\text{lin}}} \ell(\hat{\theta}^{\text{GMM}}_{u|C(S)}) - \ell(\hat{\theta}^{\text{GMM}}_{u|C(S),A(S)})$. Note that predictive exposure is $\gamma = \mathbb{E}[X]$. Thus, using standard techniques to integrate the tail, we can convert the high-probability bound into a bound on the expectation, yielding the final predictive exposure bound. Let $\mathcal{G}_1 = \mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1})$, $\mathcal{G}_2 = \mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1-\Sigma_2})$, $\mathcal{D}_1 = \mathcal{D}_{\Sigma_1}(\mathcal{U}_{\text{lin}})$, and $\mathcal{D}_2 = \mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\text{lin}})$. We know

$$\Pr\left[X > \frac{H}{n} \cdot \left(\mathcal{G}_1 + \sqrt{2\log(2/\delta)\cdot\mathcal{D}_1}\right) \cdot \left(\mathcal{G}_2 + \sqrt{2\log(2/\delta)\cdot\mathcal{D}_2}\right)\right] \le \delta.$$

Let $u = \log(2/\delta)$, then

$$\begin{aligned}
\Pr\left[X > \frac{H}{n} \cdot \left(\mathcal{G}_1 + \sqrt{2\mathcal{D}_1 u}\right) \cdot \left(\mathcal{G}_2 + \sqrt{2\mathcal{D}_2 u}\right)\right] &\le 2e^{-u} \\
\Pr[X > f(u)] &\le 2e^{-u}
\end{aligned}$$

Let $t_0 = f(\log(2))$. Thus,

$$\begin{aligned}
\mathbb{E}[X] &\le \int_0^{t_0} 1dt + \int_{t_0}^{\infty} \Pr[X > t] \\
&= t_0 + \int_{t_0}^{\infty} \Pr[X > t] \\
&= f(\log(2)) + \int_{\log(2)}^{\infty} \Pr[X > f(u)] f'(u) du \\
&= f(\log(2)) + \int_{\log(2)}^{\infty} 2e^{-u} f'(u) du \\
&= f(\log(2)) - f(\log(2)) + \int_{\log(2)}^{\infty} 2e^{-u} f(u) du \\
&= \int_{\log(2)}^{\infty} 2e^{-u} f(u) du \\
&= \int_0^{\infty} e^{-u'} f(u' + \log(2)) du && (u' = u - \log(2)) \\
&= \mathbb{E}[f(\beta + \log(2))] && (\beta \sim \text{Exp}(1)) \\
&= \frac{H}{n} \cdot \left(\mathcal{G}_1\mathcal{G}_2 + \mathcal{G}_1 \mathbb{E}\left[\sqrt{2\mathcal{D}_2(\beta + \log(2))}\right] + \mathcal{G}_2 \mathbb{E}\left[\sqrt{2\mathcal{D}_1(\beta + \log(2))}\right] + 2\sqrt{\mathcal{D}_1\mathcal{D}_2} \cdot \mathbb{E}\left[\beta + \log(2))\right]\right) \\
&= \frac{H}{n} \cdot \left(\mathcal{G}_1\mathcal{G}_2 + \mathcal{G}_1\sqrt{2\mathcal{D}_2(1 + \log(2))} + \mathcal{G}_2\sqrt{2\mathcal{D}_1(1 + \log(2))}] + 2\sqrt{\mathcal{D}_1\mathcal{D}_2}(1 + \log(2))\right) \\
&&& \text{(Jensen's inequality and } \mathbb{E}[\beta] = 1) \\
&= \frac{H}{n} \cdot \left(\mathcal{G}_1 + \sqrt{2\mathcal{D}_1(1 + \log(2))}\right) \cdot \left(\mathcal{G}_2 + \sqrt{2\mathcal{D}_2(1 + \log(2))}\right) \\
&= \frac{H}{(1-\alpha)\alpha N} \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1}) + \sqrt{2(1 + \log(2)) \cdot \mathcal{D}_{\Sigma_1}(\mathcal{U}_{\text{lin}})}\right) \left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1-\Sigma_2}) + \sqrt{2(1 + \log(2)) \cdot \mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\text{lin}})}\right).
\end{aligned}$$

The exact form of Σ_1 and Σ_2 are described in Theorem 24. □

Theorem 24. *Applying GMM with moment conditions*

$$f_{C(S)}(x,\theta) = [x-\mu] \text{ and } f_{C(S),A(S)}(x,\theta) = \begin{bmatrix} x-\mu \\ g(x,\lambda) \\ \tilde{\lambda}-\lambda \end{bmatrix},$$

we have $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu) \sim \mathcal{N}(0,\Sigma_1)$, $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)} - \mu) \sim \mathcal{N}(0,\Sigma_2)$, *where*

$$\Sigma_1 = \text{Var}[x],$$
$$\Sigma_2 = \text{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] + c_0(1,x,g,\Delta),$$
$$\Sigma_1 - \Sigma_2 = \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] - c_0(1,x,g,\Delta)$$

where $c_0(1,x,g,\Delta) = \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\Big(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]^\top \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] + \text{Var}[\Delta]^{-1}\Big)^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]^\top \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]$

Proof. First, we establish the variance of the optimal predictor $\hat{\theta}^{\text{GMM}}_{x|C(S)}$ where $\theta = (\mu)$ and

$$f_{C(S)}(x,\theta) = [x-\mu].$$

We need to compute $[G^\top\Omega^{-1}G]^{-1}$.

$$\begin{aligned} G &= \mathbb{E}\left[\tfrac{\partial f(x,\theta)}{\partial\theta}\right] \\ &= [-I_d] \end{aligned}$$

$$\begin{aligned} \Omega^{-1} &= \mathbb{E}\left[f(x,\theta)f(x,\theta)^\top\right]^{-1} \\ &= \mathbb{E}\left[(x-\mu)^2\right]^{-1} \\ &= \text{Var}[x]^{-1} \end{aligned}$$

$$\begin{aligned} [G(\theta_0)^\top\Omega^{-1}(\theta_0)G(\theta_0)]^{-1} &= [I_d\text{Var}[x]^{-1}I_d]^{-1} \\ &= \text{Var}[q(x)] \end{aligned}$$

Thus, $\Sigma_1 = \text{Var}[q(x)]$.

Next, we compute the variance of the optimal predictor $\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)}$ where $\theta = (\mu,\lambda)$ and

$$f_{C(S),A(S)}(x,\theta) = \begin{bmatrix} x-\mu \\ g(x,\lambda) \\ \tilde{\lambda}-\lambda \end{bmatrix}.$$

We need to compute $[G^\top\Omega^{-1}G]^{-1}$.

$$\begin{aligned} G &= \mathbb{E}\left[\tfrac{\partial f(x,\theta)}{\partial\theta}\right] \\ &= \begin{bmatrix} -I_d & 0 \\ 0 & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ 0 & -I_d \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \Omega^{-1} &= \mathbb{E}\left[f(x,\theta)f(x,\theta)^\top\right]^{-1} \\ &= \mathbb{E}\left[\begin{bmatrix} (x-\mu)^2 & (x-\mu)g(x,\lambda)^\top & (x-\mu)(\tilde{\lambda}-\lambda)^\top \\ g(x,\lambda)(x-\mu)^\top & g(x,\lambda)g(x,\lambda)^\top & g(x,\lambda)(\tilde{\lambda}-\lambda)^\top \\ (\tilde{\lambda}-\lambda)(x-\mu)^\top & (\tilde{\lambda}-\lambda)g(x,\lambda)^\top & (\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top \end{bmatrix}\right]^{-1} \end{aligned}$$

$$= \begin{bmatrix} \mathrm{Var}[x] & \mathbb{E}[(x-\mu)g(x,\lambda)^\top] & 0 \\ \mathbb{E}[g(x,\lambda)(x-\mu)^\top] & \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top] \end{bmatrix}^{-1}$$

Let $U^{-1} = \begin{bmatrix} \mathrm{Var}[x] & \mathbb{E}[(x-\mu)g(x,\lambda)^\top] \\ \mathbb{E}[g(x,\lambda)(x-\mu)^\top] & \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top] \end{bmatrix}^{-1}$.

$$\Omega^{-1} = \begin{bmatrix} U^{-1}[1,1] & U^{-1}[1,2] & 0 \\ U^{-1}[2,1] & U^{-1}[2,2] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix}$$

Putting it together,

$$\begin{aligned} G^\top\Omega^{-1}G &= \begin{bmatrix} -I_d & 0 & 0 \\ 0 & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] & -I_d \end{bmatrix} \begin{bmatrix} U^{-1}[1,1] & U^{-1}[1,2] & 0 \\ U^{-1}[2,1] & U^{-1}[2,2] & 0 \\ 0 & 0 & \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \begin{bmatrix} -I_d & 0 \\ 0 & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ 0 & -I_d \end{bmatrix} \\ &= \begin{bmatrix} -I_d & 0 & 0 \\ 0 & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] & -I_d \end{bmatrix} \begin{bmatrix} -U^{-1}[1,1] & U^{-1}[1,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ -U^{-1}[2,1] & U^{-1}[2,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ 0 & -\mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \\ &= \begin{bmatrix} U^{-1}[1,1] & -U^{-1}[1,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ -\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}]U^{-1}[2,1] & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}]U^{-1}[2,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] + \mathbb{E}[(\tilde{\lambda}-\lambda)(\tilde{\lambda}-\lambda)^\top]^{-1} \end{bmatrix} \\ &= \begin{bmatrix} U^{-1}[1,1] & -U^{-1}[1,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] \\ -\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}]U^{-1}[2,1] & \mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}]U^{-1}[2,2]\mathbb{E}[\frac{\partial g(x,\lambda)}{\partial\lambda}] + \mathrm{Var}[\Delta]^{-1} \end{bmatrix} \end{aligned}$$

Using Schur's complement, we can compute the entries of U^{-1}. Let $U = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, then

$$\begin{aligned} U^{-1}[1,1] &= (A-BD^{-1}C)^{-1} = [\mathrm{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]]^{-1} \\ U^{-1}[1,2] &= -(A-BD^{-1}C)^{-1}BD^{-1} = -U^{-1}[1,1]\cdot\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1} \\ U^{-1}[2,1] &= -D^{-1}C(A-BD^{-1}C)^{-1} = -U^{-1}[1,1]\cdot\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] \\ U^{-1}[2,2] &= D^{-1} + D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \\ &= \mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1} \\ &\quad + U^{-1}[1,1]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}. \end{aligned}$$

The $[1,1]$ entry of $[G^\top\Omega^{-1}G]^{-1}$ characterizes the variance of $\hat{p}$. For a matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, $M^{-1}[1,1] = A^{-1} + A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1}$, using Schur's complement.

$$\begin{aligned} [G^\top\Omega^{-1}G]^{-1}[1,1] &= [U^{-1}[1,1]]^{-1} + U^{-1}[1,1]^{-1}U^{-1}[1,2]\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\Bigg(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]U^{-1}[2,2]\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] \\ &\quad + \mathrm{Var}[\Delta]^{-1} - \mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]U^{-1}[2,1][U^{-1}[1,1]]^{-1}U^{-1}[1,2]\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\Bigg)^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] \\ &\quad \cdot U^{-1}[2,1][U^{-1}[1,1]]^{-1} \\ &= [U^{-1}[1,1]]^{-1} + [U^{-1}[1,1]]^{-1}U^{-1}[1,2]\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] \\ &\quad \cdot\left(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right] + \mathrm{Var}[\Delta]^{-1}\right)^{-1} \\ &\quad \cdot\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial\lambda}\right]U^{-1}[2,1][U^{-1}[1,1]]^{-1} \end{aligned}$$

$$
\begin{aligned}
&= [U^{-1}[1,1]]^{-1} + \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] \\
&\quad \cdot \left(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] + \mathrm{Var}[\Delta]^{-1}\right)^{-1} \\
&\quad \cdot \mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] \\
&= [U^{-1}[1,1]]^{-1} + c_0(1,x,g,\Delta) \\
&= \mathrm{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] + c_0(1,x,g,\Delta)
\end{aligned}
$$

where

$$
\begin{aligned}
c_0(1,x,g,\Delta) &= \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] \\
&\quad \cdot \left(\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right] + \mathrm{Var}[\Delta]^{-1}\right)^{-1} \\
&\quad \cdot \mathbb{E}\left[\frac{\partial g(x,\lambda)}{\partial \lambda}\right]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top].
\end{aligned}
$$

Thus, $\Sigma_2 = \mathrm{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)] + c_0(1,x,g,\Delta)$.

Finally, we have $\Sigma_1 - \Sigma_2 = \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)] - c_0(1,x,g,\Delta)$ □

Theorem 25. *Let $\xi_\Sigma \sim \mathcal{N}(0,\Sigma)$. With probability at least $1-\delta$,*

$$
\sup_{u\in K}\langle u, \xi_\Sigma\rangle \le \mathcal{G}(\Sigma^{1/2}K) + \sqrt{2\log(1/\delta)\cdot \sup_{u\in K} u^\top \Sigma u}
$$

Proof. Start by considering $\mathcal{G}_\Sigma(K)$,

$$
\mathcal{G}_\Sigma(K) = \mathbb{E}\left[\sup_{u\in K}\langle u, \xi_\Sigma\rangle\right].
$$

Let $\xi \sim \mathcal{N}(0, I_d)$. Note that $\xi_\Sigma = \Sigma^{1/2}\xi$. Thus,

$$
\begin{aligned}
\mathcal{G}_\Sigma(K) &= \mathbb{E}\left[\sup_{u\in K}\langle u, \xi_\Sigma\rangle\right] \\
&= \mathbb{E}\left[\sup_{u\in K}\langle u, \Sigma^{1/2}\xi\rangle\right] \\
&= \mathbb{E}\left[\sup_{u\in K}\langle \Sigma^{1/2}u, \xi\rangle\right] && \text{(since } \Sigma^{1/2} \text{ is symmetric)} \\
&= \mathbb{E}\left[\sup_{u'\in K_{\Sigma^{1/2}}}\langle u', \xi\rangle\right] && (K_{\Sigma^{1/2}} = \{\Sigma^{1/2}u \mid u \in K\}) \\
&= \mathcal{G}(K_{\Sigma^{1/2}})
\end{aligned}
$$

Let $\xi \sim \mathcal{N}(0, I_d)$. The mapping $g(\xi) : \xi \mapsto \sup_{u\in K}\langle \Sigma^{1/2}u, \xi\rangle$ is Lipschitz with $L = \sup_{u\in K}\sqrt{u^\top \Sigma u}$. Thus, we can use the concentration for Lipschitz functionals of standard Gaussians (Borel-CIS inequality) to get a high probability bound.

Let $Z = \sup_{u\in K}\langle u, \xi_\Sigma\rangle$. Then

$$
\Pr[Z \ge \mathbb{E}[Z] + t] = \Pr\big[Z \ge \mathcal{G}(\Sigma^{1/2}K) + t\big] \le e^{-t^2/2L^2}.
$$

Let

$$
\delta = e^{-t^2/2L^2}
$$

$$\log(1/\delta) = \frac{t^2}{2L^2}$$
$$2L^2 \log(1/\delta) = t^2$$
$$\sqrt{2L^2 \log(1/\delta)} = t$$

Thus, with probability at least $1-\delta$,

$$\sup_{u \in K} \langle u, \xi_\Sigma \rangle \le \mathcal{G}(K_{\Sigma^{1/2}}) + \sqrt{2\log(1/\delta) \cdot \sup_{u \in K} u^\top \Sigma u}.$$

□

We can use standard techniques to compute Gaussian complexity $\mathcal{G}(\Sigma^{1/2}\mathcal{Q}_{\text{lin}})$. For example, if $\mathcal{Q}_{\text{lin}}$ is the Euclidean ball of radius B, then $\mathcal{G}(\Sigma^{1/2}\mathcal{Q}_{\text{lin}}) = B\sqrt{\text{tr}(\Sigma)}$. If $\mathcal{Q}_{\text{lin}}$ is the L1 ball of radius B, then $\mathcal{G}(\Sigma^{1/2}\mathcal{Q}_{\text{lin}}) = B\sqrt{\log d} \cdot \max_i \sqrt{\Sigma[i,i]}$. We refer the reader to Bartlett & Mendelson (2003); Kakade et al. (2008) for more details.

Theorem 10. *If $\mathcal{U}_{\text{lin}}$ is contained within the Euclidean ball of radius B, for a convex, H-smooth ℓ,* $\gamma(\mathcal{P}_{\text{RS}}, \mathcal{Q}_{\text{lin}}, \ell, A) \le \frac{4HB(2+\log(2))}{(1-\alpha)\alpha N} \cdot \text{tr}(\text{Var}[x]) \cdot \sqrt{\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\text{tr}(\text{Var}[x])}}$.

Proof. From Theorem 23, we know that

$$\gamma \le \frac{H}{(1-\alpha)\alpha N}\left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1}) + \sqrt{2(1+\log(2)) \cdot \mathcal{D}_{\Sigma_1}(\mathcal{U}_{\text{lin}})}\right)\left(\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma_1-\Sigma_2}) + \sqrt{2(1+\log(2)) \cdot \mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\text{lin}})}\right)$$

When $\mathcal{U}_{\text{lin}}$ is contained within the Euclidean ball of radius B,

$$\mathcal{G}(\mathcal{U}_{\text{lin},\Sigma}) = B\sqrt{\text{tr}(\Sigma)},$$
$$\mathcal{D}_\Sigma(\mathcal{U}_{\text{lin}}) = B^2 \cdot \lambda_{\max}(\Sigma).$$

From Theorem 24,

$$\Sigma_1 = \text{Var}[x],$$
$$\Sigma_2 = \text{Var}[x] - \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] + c_0(1,x,g,\Delta),$$
$$\Sigma_1 - \Sigma_2 = \mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] - c_0(1,x,g,\Delta)$$

Then

$$\lambda_{\max}(\Sigma_1) = \lambda_{\max}(\text{Var}[x])$$

$$\begin{aligned}
\lambda_{\max}(\Sigma_1 - \Sigma_2) &= \lambda_{\max}(\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] - c_0(1,x,g,\Delta)) \\
&\le \lambda_{\max}(\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) - \lambda_{\min}(c_0(1,x,g,\Delta)) \\
&= \lambda_{\max}(\text{Var}[x]\text{Var}[x]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) \\
&\quad - \lambda_{\min}(c_0(1,x,g,\Delta)) \\
&\le \lambda_{\max}(\text{Var}[x]) \cdot \lambda_{\max}(\text{Var}[x]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) \\
&\quad - \lambda_{\min}(c_0(1,x,g,\Delta)) \\
&= \lambda_{\max}(\Sigma_1) \cdot \rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \lambda_{\min}(c_0(1,x,g,\Delta)) \\
&= \lambda_{\max}(\Sigma_1) \cdot \left(\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\lambda_{\max}(\Sigma_1)}\right) \\
&= \lambda_{\max}(\Sigma_1) \cdot \left(\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\text{tr}(\Sigma_1)}\right)
\end{aligned}$$

Next, consider

$$\text{tr}(\Sigma_1 - \Sigma_2) = \text{tr}(\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top] - c_0(1,x,g,\Delta))$$

$$
\begin{aligned}
&= \mathrm{tr}(\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) - \mathrm{tr}(c_0(1,x,g,\Delta)) \\
&= \mathrm{tr}(\mathrm{Var}[x]\mathrm{Var}[x]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) - \mathrm{tr}(c_0(1,x,g,\Delta)) \\
&\le \mathrm{tr}(\mathrm{Var}[x]) \cdot \lambda_{\max}(\mathrm{Var}[x]^{-1}\mathbb{E}[(x-\mu)g(x,\lambda)^\top]\mathbb{E}[g(x,\lambda)g(x,\lambda)^\top]^{-1}\mathbb{E}[g(x,\lambda)(x-\mu)^\top]) - \mathrm{tr}(c_0(1,x,g,\Delta)) \\
&\qquad\qquad \text{(Holder's inequality for matrices)} \\
&= \mathrm{tr}(\Sigma_1) \cdot \rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \mathrm{tr}(c_0(1,x,g,\Delta)) \\
&= \mathrm{tr}(\Sigma_1) \cdot \rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \lambda_{\min}(c_0(1,x,g,\Delta)) \\
&= \mathrm{tr}(\Sigma_1) \cdot \left(\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}\right)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\gamma &= \frac{H}{n} \cdot \left(\mathcal{G}(\mathcal{U}_{\mathrm{lin},\Sigma_1}) + \sqrt{2(1+\log(2))\mathcal{D}_{\Sigma_1}(\mathcal{U}_{\mathrm{lin}})}\right) \cdot \left(\mathcal{G}(\mathcal{U}_{\mathrm{lin},\Sigma_1-\Sigma_2}) + \sqrt{2(1+\log(2))\mathcal{D}_{\Sigma_1-\Sigma_2}(\mathcal{U}_{\mathrm{lin}})}\right) \\
&\le \frac{H}{n} \cdot \left(B\sqrt{\mathrm{tr}(\Sigma_1)} + B\sqrt{2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1)}\right)\left(B\sqrt{\mathrm{tr}(\Sigma_1-\Sigma_2)} + B\sqrt{2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1-\Sigma_2)}\right) \\
&\le \frac{HB}{n} \cdot \left(\sqrt{\mathrm{tr}(\Sigma_1)} + \sqrt{2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1)}\right) \cdot \left(\sqrt{\mathrm{tr}(\Sigma_1) \cdot \left(\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}\right)}\right. \\
&\qquad \left. + \sqrt{2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1) \cdot \left(\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}\right)}\right) \\
&= \frac{HB}{n} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}} \cdot \left(\sqrt{\mathrm{tr}(\Sigma_1)} + \sqrt{2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1)}\right)^2 \\
&= \frac{2HB}{n} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}} \cdot (\mathrm{tr}(\Sigma_1) + 2(1+\log(2)) \cdot \lambda_{\max}(\Sigma_1)) \\
&\le \frac{2HB}{n} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}} (\mathrm{tr}(\Sigma_1) + 2(1+\log(2)) \cdot \mathrm{tr}(\Sigma_1)) \\
&\le \frac{4HB}{n} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}} ((2+\log(2)) \cdot \mathrm{tr}(\Sigma_1)) \\
&= \frac{4HB(2+\log(2)) \cdot \mathrm{tr}(\Sigma_1)}{n} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\Sigma_1)}} \\
&= \frac{4HB(2+\log(2)) \cdot \mathrm{tr}(\mathrm{Var}[x])}{(1-\alpha)\alpha N} \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\mathrm{Var}[x])}}
\end{aligned}
$$

□

A.11.3 Theorem 11, Lipschitz Queries

Theorem 11. *Let query family* $\mathcal{Q}_{\mathrm{L}} = \{f(x) \mid f \text{ is L-Lipschitz}\}$. *For a convex, H-smooth loss* ℓ, $\gamma(\mathcal{P}_{\mathrm{RS}}, \mathcal{Q}_{\mathrm{L}}, \ell, A) \le \frac{4HL^2(2+\log(2))}{(1-\alpha)\alpha N} \cdot \mathrm{tr}(\mathrm{Var}[x]) \cdot \sqrt{\rho^1_{\mathrm{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\mathrm{tr}(\mathrm{Var}[x])}}$.

Proof. We want to bound $\sup_{f\in\mathcal{Q}} \ell(\hat{\theta}^{\mathrm{GMM}}_{f|C(S)}) - \ell(\hat{\theta}^{\mathrm{GMM}}_{f|C(S),A(S)})$ where ℓ is the expected log loss.

$$
\begin{aligned}
\sup_{f\in\mathcal{Q}} \ell(\hat{\theta}^{\mathrm{GMM}}_{f|C(S)}) - \ell(\hat{\theta}^{\mathrm{GMM}}_{f|C(S),A(S)}) &\le \sup_{f\in\mathcal{Q}} H \cdot |\hat{\theta}^{\mathrm{GMM}}_{f|C(S)} - \theta^f_0| \cdot |\hat{\theta}^{\mathrm{GMM}}_{f|C(S)}) - \hat{\theta}^{\mathrm{GMM}}_{f|C(S),A(S)}| \\
&= \sup_{f\in\mathcal{Q}} H \cdot |f(\hat{\theta}^{\mathrm{GMM}}_{x|C(S)}) - f(\mu^x)| \cdot |f(\hat{\theta}^{\mathrm{GMM}}_{x|C(S)}) - f(\hat{\theta}^{\mathrm{GMM}}_{x|C(S),A(S)})| \\
&= HL^2 \cdot \left\|\hat{\theta}^{\mathrm{GMM}}_{x|C(S)} - \mu\right\| \cdot \left\|\hat{\theta}^{\mathrm{GMM}}_{x|C(S)} - \hat{\theta}^{\mathrm{GMM}}_{x|C(S),A(S)}\right\|
\end{aligned}
$$

Let $n = (1-\alpha)\alpha n$. From GMM and Theorem 3, we know that $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu) \sim \mathcal{N}(0, \Sigma_1)$, $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S),A(S)} - \mu) \sim \mathcal{N}(0, \Sigma_2)$, and $\sqrt{n}(\hat{\theta}^{\text{GMM}}_{x|C(S)} - \hat{\theta}^{\text{GMM}}_{x|C(S),A(S)}) \sim \mathcal{N}(0, \Sigma_1 - \Sigma_2)$. Thus, we can use standard concentration bounds for Gaussian random variables.

With probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{Q}} \ell(\hat{\theta}^{\text{GMM}}_{f|C(S)}) - \ell(\hat{\theta}^{\text{GMM}}_{f|C(S),A(S)}) \\
&\qquad \le HL^2 \cdot \left\|\hat{\theta}^{\text{GMM}}_{x|C(S)} - \mu\right\| \cdot \left\|\hat{\theta}^{\text{GMM}}_{x|C(S)} - \hat{\theta}^{\text{GMM}}_{x|C(S),A(S)}\right\| \\
&\qquad \le \frac{HL^2}{n} \cdot \left(\sqrt{\text{tr}(\Sigma_1)} + \sqrt{2\lambda_{\max}(\Sigma_1)\log(2/\delta)}\right) \cdot \left(\sqrt{\text{tr}(\Sigma_1 - \Sigma_2)} + \sqrt{2\lambda_{\max}(\Sigma_1 - \Sigma_2)\log(2/\delta)}\right)
\end{aligned}$$

Let $X = \sup_{f\in\mathcal{Q}} \ell(\hat{\theta}^{\text{GMM}}_{f|C(S)}) - \ell(\hat{\theta}^{\text{GMM}}_{f|C(S),A(S)})$. Note that the predictive exposure is $\gamma = \mathbb{E}[X]$. Thus, using standard techniques to integrate the tail, we can convert the high-probability bound into a bound on the expectation, yielding the final predictive exposure bound. Following the proof in Theorem 23,

$$\gamma = \mathbb{E}[X] \le \frac{HL^2}{n} \cdot \left(\sqrt{\text{tr}(\Sigma_1)} + \sqrt{2\lambda_{\max}(\Sigma_1)(1+\log(2))}\right) \cdot \left(\sqrt{\text{tr}(\Sigma_1 - \Sigma_2)} + \sqrt{2\lambda_{\max}(\Sigma_1 - \Sigma_2)(1+\log(2))}\right)$$

We can further simplify this by following Theorem 10 and substituting in Σ_1 and Σ_2 from Theorem 24

$$\gamma = \frac{4HL^2(2+\log(2)) \cdot \text{tr}(\Sigma_1)}{(1-\alpha)\alpha N} \cdot \sqrt{\rho^1_{\text{cc}}(x, g(x,\lambda))^2 - \frac{\lambda_{\min}(c_0(1,x,g,\Delta))}{\text{tr}(\Sigma_1)}}$$

□

A.12 Helpful Theorems

Theorem 26 (McFadden (1999) Chapter 3, Lemma 3.1)**.** *$(G^\top W G)^{-1} G^\top W \Omega W G (G^\top W G)^{-1} - (G^\top \Omega^{-1} G)^{-1}$ is positive semidefinite.*

Theorem 27 (McFadden (1999) Chapter 3, Lemma 3.2)**.** *If $Y_n \to_d Y_0$ and f is a continuous function of an open set that contains the support of Y_0, then*

$$f(Y_n) \to_d f(Y_0).$$

Theorem 28 (McFadden (1999) Chapter 3, Lemma 3.4)**.** *If $Y_n(\theta) \to_p Y_0(\theta)$ uniformly for $\theta \in \Theta \subseteq \mathbb{R}^d$, random vectors $\tau_0, \tau_n \in \Theta$ satisfy $\tau_n \to_p \tau_0$, and $Y_0(\theta)$ is almost surely continuous at τ_0, then*

$$Y_n(\tau_n) \to_p Y_0(\tau_0).$$