Auditing Privacy of Additive Noise Mechanisms Using Linear Predictive Models

Monica Welfert*, Nathan Stromberg*, Mario Diaz†, James Melbourne‡ and Lalitha Sankar*
*Arizona State University, USA, †IIMAS, Mexico, ‡CIMAT, Mexico
Email: {mwelfert, nstrombe, lsankar}@asu.edu, mario.diaz@sigma.iimas.unam.mx, james.melbourne@cimat.mx

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. We propose a computationally feasible privacy auditing framework using minimum mean-squared error (MMSE) estimation and linear auditing models. Our approach provides theoretical lower bounds on the true MMSE of inferring sensitive features from noisy observations of other correlated features. The bounds are in terms of the empirical MMSE under a restricted hypothesis class and a decomposable error term capturing finite sample and approximation effects. For linear auditing models, we derive closed-form bounds for classes of relationships between the private and non-private features, including linear mappings, binary symmetric channels, and class-conditional Gaussian models. Through empirical evaluation, we demonstrate that our linear model-based auditing framework serves as a powerful yet tractable tool for MMSE-based privacy auditing that balances theoretical guarantees with practical efficiency.**

I. Introduction

Evaluating the minimum mean-squared error (MMSE) in estimating one random variable from another is a classical problem in communications and machine learning. This classical problem is very relevant in privacy applications, where it provides a framework for quantifying information leakage: by measuring how accurately a sensitive variable S can be estimated from a feature vector X, we can establish privacy guarantees for data sharing.

To guarantee privacy, practitioners often release noised versions X^σ of raw features X to deliberately obfuscate sensitive features S. Our work introduces a novel approach to auditing these privacy-preserving noise mechanisms, specifically additive mechanisms, using simple linear models. There exist very rigorous but computationally expensive approaches for auditing differentially private models (see [1] and references therein). Our goal in this work is to look at a somewhat simpler machine learning adversary and provide average-case guarantees that are very typical of machine learning predictive methods. This may not guarantee differential privacy, but these are very relevant in fairness settings, where some variables cannot be released, or also in restrictive privacy settings, where we are worried about sensitive features rather than individuals.

Previous work [2] has attempted similar guarantees using neural networks, but such approaches require computing or bounding Barron's constant [3]—a generally infeasible task. Our simpler auditing models offer key advantages: they avoid the computational challenges of neural networks while maintaining meaningful privacy guarantees, and they address the challenge that the bounds require learning an auditing model with non-zero training MSE. This is particularly important as neural networks' non-linearity makes them prone to overfitting, even on complex datasets.

Our primary contribution is establishing lower bounds on the true MMSE of S given the noisy observed X^σ of the following form:

$$\operatorname{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_n^{\mathcal{H}} \leq \operatorname{mmse}(S|X^\sigma), \tag{1}$$

where $\operatorname{mmse}_n^{\mathcal{H}}(S|X^\sigma)$ is the empirical MMSE given n samples and a hypothesis class $\mathcal{H}$ and $\epsilon_n^{\mathcal{H}}$ is a positive number depending on n and the parameters of $\mathcal{H}$. We focus on specific classes of relationships between S and X and demonstrate the following:

- We decompose $\epsilon_n^{\mathcal{H}}$ into two components: (i) a finite sample term dependent on n, bounded using large deviation theory, and (ii) an approximation error term dependent on $\mathcal{H}$ parameters (Proposition 1).
- When $\mathcal{H}$ is taken to be the hypothesis class of the sigmoid function composed with a linear function, we obtain a closed-form bound on the approximation error in terms of the statistics of X^σ and the optimizer of the true $\operatorname{mmse}(S|X^\sigma)$ (Proposition 2).
- Under Gaussian noise and binary S, we explicitly compute the bound on the approximation error for the following relationships between S and X: (i) linear relationship (Proposition 3), (ii) class-conditional Gaussian with equal covariances (Proposition 3), (iii) binary symmetric channel (Theorem 1), and (iv) class-conditional Gaussian with different covariances (Theorem 2 and Corollary 1).

II. Problem Setup

Given jointly distributed random variables $X \in \mathbb{R}^d$ and $S \in \mathbb{R}$, the minimum mean square error in estimating S given X is defined as

$$\operatorname{mmse}(S|X) \coloneqq \inf_{h \text{ meas.}} \mathbb{E}\left[(S - h(X))^2\right], \tag{2}$$

where the infimum is taken over all (Borel) measurable functions $h : \mathbb{R}^d \to \mathbb{R}$. The infimum in (2) is attained by the conditional expectation of S given X, i.e.,

$$\operatorname{mmse}(S|X) = \mathbb{E}\left[(S - \eta(X))^2\right], \tag{3}$$

where $\eta(X) \overset{\text{a.s.}}{=} \mathbb{E}[S|X]$. Note that if $S \overset{\text{a.s.}}{=} h_0(X)$ for some function $h_0 : \mathbb{R}^d \to \mathbb{R}$, then $\operatorname{mmse}(S|X) = 0$. Also, note that

if X and S are independent, then the MMSE is maximal and $\mathrm{mmse}(S|X) = \mathbb{E}\left[(S - \mathbb{E}[S])^2\right]$.

In the context of estimation-theoretic privacy, Asoodeh *et al.* [4] introduced the notion of ϵ-*weak estimation privacy* to denote that

$$\mathrm{mmse}(S|X) \geq (1-\epsilon)\mathbb{E}\left[(S - \mathbb{E}[S])^2\right]. \tag{4}$$

Based on (2), $\mathrm{mmse}(S|X)$ quantifies the ability of an adversary to approximate S, in the expected square-loss sense, upon observing X. Since a larger MMSE amounts to better privacy, estimation-theoretic privacy guarantees make natural sense as lower bounds for $\mathrm{mmse}(S|X)$ as expressed in (4).

Also, when S is binary, the MMSE serves as a lower bound for the probability of error. Specifically, if $S \in \{\pm 1\}$, then

$$\begin{aligned} P_{\text{error}}(S|X) &= \inf_{h:\mathbb{R}^d \to \{\pm 1\}} \mathbb{E}\left[\mathbb{1}_{S \neq h(X)}\right] \\ &= \inf_{h:\mathbb{R}^d \to \{\pm 1\}} \mathbb{E}\left[\frac{(S-h(X))^2}{4}\right] \\ &\geq \frac{1}{4} \inf_{h \text{ meas.}} \mathbb{E}\left[(S-h(X))^2\right] \\ &= \frac{1}{4}\mathrm{mmse}(S|X). \end{aligned} \tag{5}$$

Thus, for binary S, any lower bound for $\mathrm{mmse}(S|X)$ gives rise to a lower bound for $P_{\text{error}}(S|X)$. This observation further illustrates the importance of studying lower bounds for the MMSE in the context of privacy, where probability of correctly guessing $(1 - P_{\text{error}})$ has also been used as an information leakage measure [5]–[7].

Assume that $X \in \mathbb{R}^d$ are usable features of an individual which are correlated with a private/sensitive attribute $S \in [0,1]$, e.g., X and S could be be height and gender, respectively. Due to privacy concerns, a data analyst might not be able to observe X but a *sanitized* version of it. In precise terms, we have a Markov chain $S - X - X^\sigma$ where S represents sensitive information (e.g., gender), X represents non-sensitive information (e.g., height), and X^σ is a noisy version of X. In this work, we focus on the additive Gaussian mechanism, a popular sanitization method in the information-theoretic and the differential privacy literature, e.g., [4], [8], [9]. Given $\sigma > 0$, we define

$$X^\sigma \coloneqq X + \sigma Z, \tag{6}$$

where $Z \sim \mathcal{N}(0, I)$ is independent of X and S. Let f_i and f_i^σ be the (conditional) density of $X|S=i$ and $X^\sigma|S=i$, respectively, for $i \in \{0,1\}$. Note that $f_i^\sigma(x) = (f_i * K_\sigma)(x)$, where $*$ is the convolution operator and K_σ is the density of σZ. Then the conditional expectation of S given X^σ is given by

$$\eta^\sigma(x) = \frac{p f_1^\sigma(x)}{p f_1^\sigma(x) + \bar{p} f_0^\sigma(x)} = s(\theta^\sigma(x)), \quad x \in \mathbb{R}^d, \tag{7}$$

where $p = \mathbb{P}(S=1)$, $\bar{p} \coloneqq 1-p$, $s(z) = 1/(1+e^{-z})$ is the sigmoid function, and

$$\theta^\sigma(x) = \log\left(\frac{p f_1^\sigma(x)}{\bar{p} f_0^\sigma(x)}\right). \tag{8}$$

Let $\mathcal{H}$ be the hypothesis class associated with a collection of functions $h_{\mathcal{H}} : \mathbb{R}^d \to [0,1]$. In this work we propose the following estimator of the MMSE of S given X^σ. Given a random sample $\{(X_i^\sigma, S_i)\}_{i=1}^n$, we define

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) \coloneqq \inf_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n (S_i - h(X_i^\sigma))^2, \tag{9}$$

i.e., $\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma)$ is the minimum empirical square-loss attained by a model $h \in \mathcal{H}$. As we wish to give guarantees on the privacy provided by adding Gaussian noise to a dataset, our goal is therefore to establish a (probabilistic) lower bound of the form

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_n^{\mathcal{H}} \leq \mathrm{mmse}(S|X^\sigma), \tag{10}$$

where $\epsilon_n^{\mathcal{H}}$ is a positive number depending on the sample size n and the parameters of the hypothesis class $\mathcal{H}$.

Notation: For functions, we use $\|\cdot\|_2$ to denote the 2-norm with respect to the distribution of X^σ, i.e., $\|h\|_2^2 = \mathbb{E}[|h(X)|^2]$. We define $\bar{p} \coloneqq 1-p$.

III. Main Results

By definition, $\mathrm{mmse}(S|X^\sigma)$ represents the smallest expected square-loss achievable by *any* measurable function. This distinguishes our bound in (10) from classical statistical learning results, such as Rademacher complexity bounds, where the expected loss is minimized only over a specific hypothesis class $\mathcal{H}$. This distinction necessitates explicit consideration of the approximation error, i.e., how well functions in a hypothesis class $\mathcal{H}$ can approximate the true conditional expectation. The following result provides a lower bound for $\mathrm{mmse}(S|X^\sigma)$ that depends on three components: (i) the empirical estimator $\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma)$ as defined in (9); (ii) the approximation error incurred when estimating η^σ in (7) using functions from the chosen hypothesis class $\mathcal{H}$; and (iii) the estimation error, which accounts for the uncertainty in learning the optimal function in $\mathcal{H}$ using only a finite sample of size n.

Proposition 1. *Let $n \in \mathbb{N}$, $\mathcal{H} \subseteq \{f : \mathbb{R}^d \to [0,1]\}$ and $h_{\mathcal{H}}^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}[(S - h(X^\sigma))^2]$. If $S \in [0,1]$, then, with probability at least $1-\delta$,*

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_{n,\delta}^{\mathcal{H}} \leq \mathrm{mmse}(S|X^\sigma), \tag{11}$$

where

$$\epsilon_{n,\delta}^{\mathcal{H}} = \epsilon_C + \epsilon_A, \tag{12}$$

for

$$\epsilon_C \coloneqq \sqrt{\frac{\log(1/\delta)}{2n}} \quad \textit{and} \quad \epsilon_A \coloneqq \|\eta^\sigma - h_{\mathcal{H}}^*\|_2^2. \tag{13}$$

Here, ϵ_C denotes the error resulting from estimating $h_{\mathcal{H}}^*$ using n samples and ϵ_A denotes the error resulting from approximating η^σ by $h_{\mathcal{H}}^*$.

Proof sketch. We decompose the difference of the empirical MMSE and the true MMSE into the estimation error and approximation error by adding and subtracting the MMSE

from using a restricted hypothesis class but an infinite number of samples. We then apply large deviation bounds to bound the estimation error. Proof details are in Appendix A.

Because the sample complexity term ϵ_C in (12) only depends on the number of samples n, we focus our attention on the approximation error term ϵ_A. In order to bound this term, we consider the simple setting where $\mathcal{H}$ is the hypothesis class of the sigmoid function composed with linear functions. In the following result, we show that we can obtain a closed-form bound on ϵ_A in terms of the statistics of $\theta^\sigma(X^\sigma)$ and X^σ.

Proposition 2. *Let $\mathcal{H}_L$ denote the hypothesis class associated with linear functions $\theta_L : \mathbb{R}^d \to \mathbb{R}$ of the form*

$$\theta_L(x) = a^{\mathrm{T}}x + b, \tag{14}$$

where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then there exists $\theta_L^* \in \mathcal{H}_L$ such that

$$a^* = \mathit{Cov}(\theta^\sigma(X^\sigma), X^\sigma)\mathit{Var}(X^\sigma)^{-1}, \tag{15}$$

$$b^* = \mathbb{E}[\theta^\sigma(X^\sigma)] - (a^*)^T\mathbb{E}[X^\sigma]. \tag{16}$$

and

$$\begin{aligned}\epsilon_A \leq\ & \mathit{Var}(\theta^\sigma(X^\sigma)) \\ & - \mathit{Cov}(\theta^\sigma(X^\sigma), X^\sigma)\mathit{Var}(X^\sigma)^{-1}\mathit{Cov}(X^\sigma, \theta^\sigma(X^\sigma)).\end{aligned} \tag{17}$$

Proof sketch. Using the fact that $\eta^\sigma(x) = s(\theta^\sigma(x))$, we can write $h_{\mathcal{H}}^*(x) = s(\theta_L^*(x))$ for some $\theta_L^* \in \arg\min_{\theta_L \in \mathcal{H}_L} \mathbb{E}\left[(\theta^\sigma(X^\sigma) - \theta_L(X^\sigma))^2\right]$. We can then bound ϵ_A as follows:

$$\epsilon_A = \|s \circ \theta^\sigma - s \circ \theta_L^*\|_2^2 \leq \|\theta^\sigma - \theta_L^*\|_2^2, \tag{18}$$

where the inequality follows from the fact that s is 1-Lipschitz. Proof details are in Appendix B.

We can therefore bound ϵ_A by approximating θ^σ using the linear function θ_L^*. We now present several examples for various increasingly complex relationships between S and X to gain some intuition about the behavior of ϵ_A as a function of σ.

Proposition 3. *If $S \sim \mathrm{Ber}(p)$ and*
(i) $X = aS + b$ for some $a, b \in \mathbb{R}^d$ or
(ii) $(X|S = s) \sim N(\mu_s, \Sigma)$ for mean $\mu_s \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$,
then $\epsilon_A = 0$. With probability at least $1 - \delta$ for $n \in \mathbb{N}$ and $\sigma > 0$,

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_{n,\delta}^{\mathcal{H}} \leq \mathrm{mmse}(S|X^\sigma), \tag{19}$$

where

$$\epsilon_{n,\delta}^{\mathcal{H}} = \epsilon_C. \tag{20}$$

Proof sketch. For setting (ii), computing θ^σ in (8) yields

$$\theta^\sigma(x) = (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}x + \frac{1}{2}\left(\mu_0^{\mathrm{T}}\tilde{\Sigma}^{-1}\mu_0 - \mu_1^{\mathrm{T}}\tilde{\Sigma}^{-1}\mu_1\right),$$

where $\tilde{\Sigma} \coloneqq \Sigma + \sigma^2 I$. Since θ^σ linear in x, it can therefore be recovered exactly by θ_L^*. Moreover, since setting (i) is a special case of setting (ii), the same conclusion holds for setting (i) as well. See Appendix C for a complete proof.

Remark 1. Once Gaussian noise is added to X, the setting (i) in Proposition 3 reduces to a special case of setting (ii), for which it is known that a linear separator is optimal. Therefore the only difference between $\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma)$ and $\mathrm{mmse}(S|X^\sigma)$ is the sample complexity term ϵ_C, which can be made small given enough samples.

It is of more interest to look at settings where the optimal predictor is not linear to understand how ϵ_A affects the lower bound on the true MMSE. The simplest example we consider is a binary symmetric channel, which is very relevant in communication systems.

Theorem 1 (Binary Symmetric Channel)**.** *If $S \sim \mathit{Ber}(p)$, $N \sim \mathit{Ber}(p_N)$, and $X = S \oplus N$, then*

$$\begin{aligned}\epsilon_A &\leq \epsilon_A^{\mathit{BSC}} \\ &\coloneqq \frac{(1-2p_N)^2}{\sigma^2} - \frac{(1-2p_N)^2}{p\overline{p}(1-2p_N)^2 + p_N\overline{p_N} + \sigma^2} \\ &\quad + \frac{2(1-2p_N)^2\left(p\overline{p}(1-2p+2p^2) - p_N\overline{p_N}\right)}{\sigma^4} \\ &\quad - \frac{2(1-2p_N)\left(p^2(2p-3)(2p_N-1)^2\right)}{\sigma^2\left(p\overline{p}(1-2p_N)^2 + p_N\overline{p_N} + \sigma^2\right)} \\ &\quad - \frac{2(1-2p_N)\left(2p_N^2\overline{p_N} + p(1-6p_N+6p_N^2)\right)}{\sigma^2\left(p\overline{p}(1-2p_N)^2 + p_N\overline{p_N} + \sigma^2\right)} \\ &\quad + \mathcal{O}\left(\frac{C_1(p,p_N)(1-2p_N)^2}{\sigma^6}\right) \\ &\quad + \mathcal{O}\left(\frac{C_2(p,p_N)}{\sigma^4\left(p\overline{p}(1-2p_N)^2 + p_N\overline{p_N} + \sigma^2\right)}\right)\end{aligned} \tag{21}$$

for some constants $C_1(p, p_N)$ and $C_2(p, p_N)$. With probability at least $1 - \delta$ for $n \in \mathbb{N}$ and $\sigma > 0$,

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_{n,\delta}^{\mathcal{H}} \leq \mathrm{mmse}(S|X^\sigma), \tag{22}$$

where

$$\epsilon_{n,\delta}^{\mathcal{H}} \leq \epsilon_C + \epsilon_A^{\mathit{BSC}}. \tag{23}$$

Proof sketch. Computing θ^σ in (8) yields

$$\theta^\sigma(x) = \log\left(\frac{p}{\overline{p}}\right) + \log\left(\frac{p_N + \overline{p_N}e^{(2x-1)/2\sigma^2}}{\overline{p_N} + p_N e^{(2x-1)/2\sigma^2}}\right). \tag{24}$$

We use a series expansion for the second term in (24) in addition to the binomial theorem and the series expansion for the exponential function in order to derive approximations of the expectations in (17). Proof details are in Appendix D.

Remark 2. For large noise levels σ, the approximation error bound $\epsilon_A^{\mathrm{BSC}}$ decays as $\mathcal{O}(1/\sigma^2)$. Thus, $\mathrm{mmse}(S|X^\sigma)$ is within $\mathcal{O}(1/\sqrt{n} + 1/\sigma^2)$ of $\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma)$. With sufficient samples n, this convergence is dominated by the noise term $\mathcal{O}(1/\sigma^2)$.

This behavior has an intuitive explanation: as the noise level σ increases, the class-conditional distributions increasingly resemble pure Gaussians rather than Gaussian mixtures, causing the optimal estimator θ^σ in (24) to approach linearity and eventually become constant in the limit. Consequently, a linear function provides an increasingly good approximation. Additionally, as the class-conditional distributions overlap more substantially, both the true MMSE and the empirical estimate grow with σ, reflecting diminishing ability to recover the sensitive attribute S from the noised features X^σ. While larger σ provides stronger privacy guarantees, the choice of noise level should ultimately be dictated by the utility of the noised features X^σ. In this setting, our analysis shows that a linear model serves as an effective privacy auditing tool, providing reliable estimates of information leakage with minimal computational overhead.

Our next example examines the case where X follows a class-conditional Gaussian distribution with distinct covariance matrices. For this setting, we can derive exact closed-form expressions for each term in the bound given by (17), rather than relying on approximations.

Theorem 2 (Class Conditional Vector Gaussian)**.** *Suppose $S \sim Ber(p)$ and $(X|S=s) \sim \mathcal{N}(\mu_s, \Sigma_s)$ for $s \in \{0,1\}$, where $\mu_s \in \mathbb{R}^d$ and $\Sigma_s \in \mathbb{R}^{d\times d}$ is a positive definite covariance matrix. Let $\tilde{\Sigma}_s \coloneqq \Sigma_s + \sigma^2 I$ for $\sigma > 0$, $\mu_d \coloneqq \mu_1 - \mu_0$,*

$$A = \frac{1}{2}\left(\tilde{\Sigma}_0^{-1} - \tilde{\Sigma}_1^{-1}\right), \tag{25}$$

$$b = \tilde{\Sigma}_1^{-1}\mu_1 - \tilde{\Sigma}_0^{-1}\mu_0, \tag{26}$$

$$c = \frac{1}{2}\mu_0^{\mathrm{T}}\tilde{\Sigma}_0^{-1}\mu_0 - \frac{1}{2}\mu_1^{\mathrm{T}}\tilde{\Sigma}_1^{-1}\mu_1 + \frac{1}{2}\log\left(\frac{|\tilde{\Sigma}_0|}{|\tilde{\Sigma}_1|}\right) + \log\left(\frac{p}{\overline{p}}\right), \tag{27}$$

$$M_s \coloneqq \sum_{j=1}^{d} \lambda_j^{(s)} + b^{\mathrm{T}}\mu_s + \mu_s^{\mathrm{T}} A \mu_s + c, \tag{28}$$

$$V_s \coloneqq \sum_{j=1}^{d} 2(\lambda_j^{(s)})^2 + (u_j^{(s)})^2, \tag{29}$$

where $\lambda_j^{(s)}$, $j \in \{1,\dots,d\}$, are the eigenvalues of $\tilde{\Sigma}_y^{1/2} A \tilde{\Sigma}_y^{1/2}$ with corresponding eigenvectors as the columns of a matrix Q_y, i.e., $Q_s^{\mathrm{T}}\tilde{\Sigma}_s^{1/2} A \tilde{\Sigma}_s^{1/2} Q_s = diag(\lambda_1^{(s)},\dots,\lambda_d^{(s)})$, and

$$u^{(s)} = (u_1^{(s)},\dots,u_d^{(s)})^{\mathrm{T}} = Q_s^{\mathrm{T}}\tilde{\Sigma}_s^{1/2}(b + 2A\mu_s). \tag{30}$$

Then, for

$$VAR_1 = pV_1 + \overline{p}V_0 + p\overline{p}\left(M_1 - M_0\right)^2, \tag{31}$$

$$\begin{aligned} COV = {} & 2p\mu_1^{\mathrm{T}} A\Sigma_1 + 2\overline{p}\mu_0^{\mathrm{T}} A\Sigma_0 + 2\sigma^2\left[p\mu_1 + \overline{p}\mu_0\right]^{\mathrm{T}} A \\ & + p\overline{p}\left[tr(A(\Sigma_1 - \Sigma_0)) + \mu_1^{\mathrm{T}} A\mu_1 - \mu_0^{\mathrm{T}} A\mu_0\right]\mu_d^{\mathrm{T}} \\ & + b^{\mathrm{T}}\left[p\Sigma_1 + \overline{p}\Sigma_0 + \sigma^2 I + p\overline{p}\mu_d\mu_d^{\mathrm{T}}\right] \end{aligned} \tag{32}$$

$$VAR_2 = p\Sigma_1 + \overline{p}\Sigma_0 + p\overline{p}\mu_d\mu_d^{\mathrm{T}} + \sigma^2 I, \tag{33}$$

$$\epsilon_A \le \epsilon_A^{CCG} \coloneqq VAR_1 - COV * VAR_2^{-1} * COV^{\mathrm{T}}. \tag{34}$$

and with probability at least $1-\delta$ for $n \in \mathbb{N}$,

$$\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \epsilon_{n,\delta}^{\mathcal{H}} \le \mathrm{mmse}(S|X^\sigma), \tag{35}$$

where

$$\epsilon_{n,\delta}^{\mathcal{H}} \le \epsilon_C + \epsilon_A^{CCG}. \tag{36}$$

Proof sketch. It is well-known that the optimal separator is quadratic [10], i.e.,

$$\theta^\sigma(x) = x^{\mathrm{T}} A x + b^{\mathrm{T}} x + c, \tag{37}$$

where A and b are defined in (25) and (26), (27) respectively. We then compute each term in (17) as in (31), (32) and (33) by conditioning on S and using moments of quadratic forms of Gaussian random vectors. A complete proof can be found in Appendix E.

To gain more intuition about how the class-conditional distribution parameters μ_s, Σ_s, $s \in \{0,1\}$ and noise parameter σ affect ϵ_A, we consider the special case when $\Sigma_s = \sigma_s^2 I$.

Corollary 1. *Suppose $S \sim Ber(p)$ and $(X|S=s) \sim \mathcal{N}(\mu_s, \sigma_s^2 I)$ for $s \in \{0,1\}$, where $\mu_s \in \mathbb{R}^d$ and $\sigma_s > 0$. Then for $\sigma > 0$,*

$$\epsilon_A \le \epsilon_A^{CCG} = \frac{(\sigma_1^2 - \sigma_0^2)^2(q_1 + q_2\sigma^2 + q_3\sigma^4 + 2d\sigma^6)}{4(r_1 + r_2\sigma^2 + r_3\sigma^4 + r_4\sigma^6 + r_5\sigma^8 + \sigma^{10})}, \tag{38}$$

where

$$\begin{aligned} q_1 = {} & \|\mu_1 - \mu_0\|_2^4\left(p^2\overline{p}\sigma_1^2 + p\overline{p}^2\sigma_0^2\right) + 2d\sigma_0^6 - d^2p^3(\sigma_1^2 - \sigma_0^2)^3 \\ & + 2p\overline{p}(2+d)\sigma_0^2\sigma_1^2\|\mu_1 - \mu_0\|_2^2 + p^2\left(d(5d-2)\sigma_0^4\sigma_1^2\right. \\ & \left. -2d(2d+1)\sigma_0^2\sigma_1^4 + d(2+d)\sigma_1^6 - 2d(d-1)\sigma_0^6\right) \\ & + p\left(d(d-4)\sigma_0^6 - 2d(d-1)\sigma_0^4\sigma_1^2 + d(2+d)\sigma_0^2\sigma_1^4\right), \\ q_2 = {} & p\overline{p}\|\mu_1 - \mu_0\|_2^2\left(\|\mu_1 - \mu_0\|_2^2 + 2(2+d)(\sigma_0^2 + \sigma_1^2)\right) \\ & - p^2 d(d-4)(\sigma_1^2 - \sigma_0^2)^2 + p\left(d(d-10)\sigma_0^4\right. \\ & \left. -2d(d-4)\sigma_0^2\sigma_1^2 + d(2+d)\sigma_1^4\right) + 6d\sigma_0^4, \\ q_3 = {} & 2p\overline{p}(2+d)\|\mu_1 - \mu_0\|_2^2 + 6d(p\sigma_1^2 + \overline{p}\sigma_0^2), \\ r_1 = {} & \sigma_0^4\sigma_1^4\left(p\overline{p}\|\mu_1 - \mu_0\|_2^2 + p\sigma_1^2 + \overline{p}\sigma_0^2\right), \\ r_2 = {} & \sigma_0^2\sigma_1^2\left(2p\overline{p}(\sigma_1^2 + \sigma_0^2)\|\mu_1 - \mu_0\|_2^2 + 2p\sigma_1^4 + 2\overline{p}\sigma_0^4 + 3\sigma_0^2\sigma_1^2\right), \\ r_3 = {} & p\overline{p}(\sigma_0^4 + 4\sigma_0^2\sigma_1^2 + \sigma_1^4)\|\mu_1 - \mu_0\|_2^2 + p\sigma_1^6 + 3(1+p)\sigma_0^2\sigma_1^4 \\ & + 3(2-p)\sigma_0^4\sigma_1^2 + \overline{p}\sigma_0^6, \\ r_4 = {} & 2p\overline{p}(\sigma_1^2 + \sigma_0^2)\|\mu_1 - \mu_0\|_2^2 + (2p+1)\sigma_1^4 + (3-2p)\sigma_0^4 \\ & + 6\sigma_0^2\sigma_1^2, \\ r_5 = {} & p\overline{p}\|\mu_1 - \mu_0\|_2^2 + (3-p)\sigma_0^2 + (2+p)\sigma_1^2. \end{aligned}$$

Proof sketch. Substituting $\Sigma_s = \sigma_s^2 I$, and consequently $\tilde{\Sigma}_s = (\sigma_s^2 + \sigma^2)I$, into Theorem 2, we then simplify and manipulate the resulting expressions to obtain the result in (38). Proof details can also be found in Appendix F.

Remark 3. For large values of σ, $\epsilon_A^{\mathrm{CCG}}$ decays as $\mathcal{O}(1/\sigma^2)$.

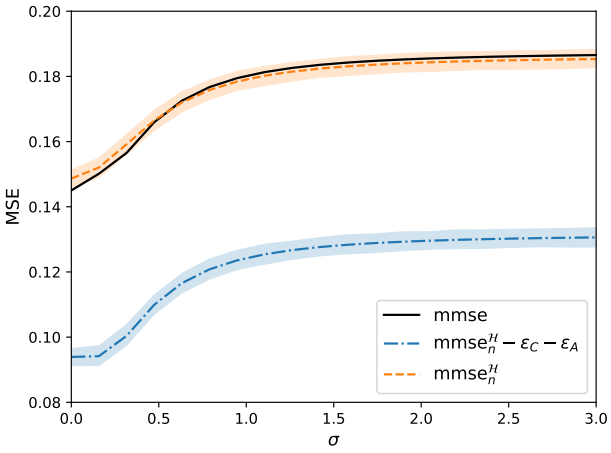

Fig. 1. Bound on the BSC MMSE as a function of the noise variance σ^2. Here MMSE and ϵ_A are estimated from 1M samples and $\mathrm{mmse}_n^{\mathcal{H}}$ is the linear training MMSE with $n = 500$ samples. Note that most of the gap between the true MMSE and the given bound comes from ϵ_C and a linear auditing model is sufficient for bounding MMSE.

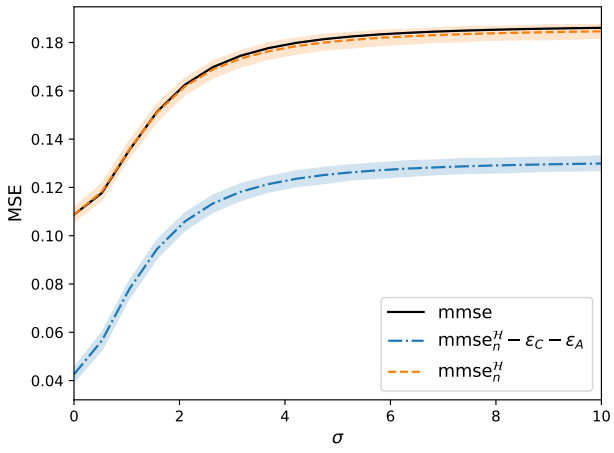

Fig. 2. Bound on the CCG MMSE using a linear auditing MMSE. Here MMSE and ϵ_A are estimated from 1M samples and $\mathrm{mmse}_n^{\mathcal{H}}$ is the linear training MMSE with $n = 500$ samples. We see that despite the linear auditing model being insufficient to fully capture the optimal model in this setting, we are still able to provide a good lower bound on the MMSE.

IV. Experimental Results

To empirically verify the lower bounds presented, we estimate several key terms through simulation. We use a million data points to obtain a good estimate of the true MMSE in (2) as well as to estimate the optimal linear model θ_L^* in (14) using the parameters in (15) and (16). Additionally, we train our auditing model as a logistic model (linear composed with sigmoid) using square loss via gradient descent; the resulting training error for a chosen number of samples n yields $\mathrm{mmse}_n^{\mathcal{H}}$. Given θ_L^* and mmse, we compute ϵ_A using the equality in (18). Taking these quantities, we demonstrate the efficacy of our lower bound below.

We first illustrate in Figure 1 the BSC setting described in Theorem 1 for parameter choices $p = 1/4$ and $p_N = 1/4$. We choose $n = 500$ for which $\epsilon_C \approx 0.05$ for a bound with probability 95%. The black solid curve is the empirical estimate of the true MMSE using a million samples of (X, S), while the orange dashed curve is the *auditing MMSE* (i.e., the empirical MMSE) $\mathrm{mmse}_n^{\mathcal{H}}$ for a linear model with sigmoid activation trained with $n = 500$ samples. The solid blue curve is the bound provided in Proposition 1 computed using the (million sample-based) empirical estimates of ϵ_A and MMSE. We note that the bound provided is quite tight and that the majority of the gap is due to ϵ_C. For very small σ, the linear auditing model is suboptimal and has a larger training MMSE than the true MMSE, as evidenced in Figure 1.

We next consider the one-dimensional class-conditional Guassian setting as described in Corollary 1 with parameters $p = 1/4, d = 1, \sigma_0 = 1, \sigma_1 = 3, \mu_0 = -1, \mu_1 = 1$. While a linear model is not optimal in this setting for small σ, we are still able to effectively bound the true MMSE for every noise level. Note that the loosest bound is at $\sigma = 0$, where the approximation error is the largest, following the results of Corollary 1. Despite this, we see that a linear auditor provides a meaningful bound for reasonable values of σ^2.

V. Concluding Remarks

In practical applications, we face two major challenges when working with the bound on ϵ_A in (17): (i) the conditional densities f_0 and f_1 may be unknown, making it impossible to derive θ^σ explicitly; and (ii) even when these densities are known, they might result in θ^σ being too complex to compute the terms in the bound directly. To address these challenges, we can further bound (17) using Poincaré inequalities. These kinds of inequalities have previously been used in communication systems to bound the true MMSE [11], but have not been effectively applied for the setting with a finite number of samples. We view the effort in this paper as a first *verifiable* step towards obtaining such general (albeit weaker) bounds. Thus, we focus here on simple data distributions where closed form computations are feasible.

More broadly, the lower bounds we have presented on the true MMSE are also relevant to communication systems (S and X are the transmitted bit and signal, respectively, over a Gaussian channel) where MMSE continues to be a key measure of goodness. In particular, as these systems employ complex non-linear encoders and decoders, lower bounds on the true MMSE can provide a simple means of evaluating the complex models for the finite blocklength setting.

Acknowledgment

This work is supported in part by NSF grants CIF-1901243, CIF-2007688, CIF-2312666, and SCH-2205080. We lost one of our collaborators, Prof. Mario Diaz, halfway through this effort but he continues to be the inspiration for this work.

REFERENCES

[1] K. Pillutla, G. Andrew, P. Kairouz, H. B. McMahan, A. Oprea, and S. Oh, "Unleashing the power of randomization in auditing differentially private ml," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 66 201–66 238, 2023.

[2] M. Diaz, P. Kairouz, J. Liao, and L. Sankar, "Neural network-based estimation of the MMSE," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1023–1028.

[3] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[4] S. Asoodeh, F. Alajaji, and T. Linder, "Privacy-aware MMSE estimation," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 1989–1993.

[5] C. Braun, K. Chatzikokolakis, and C. Palamidessi, "Quantitative notions of leakage for one-try attacks," *Electronic Notes in Theoretical Computer Science*, vol. 249, pp. 75–91, 2009.

[6] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2019.

[7] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 1949–1978, 2019.

[8] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[9] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, p. 15, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009.

[11] I. Zieder, A. Dytso, and M. Cardone, "An mmse lower bound via poincare inequality," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 957–962.

[12] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[13] A. Mathai and S. Provost, *Quadratic Forms in Random Variables*, ser. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1992.

APPENDIX

A. Proof of Proposition 1

For ease of notation, we define

$$\Delta \coloneqq \mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \mathrm{mmse}(S|X^\sigma). \tag{39}$$

Note that we can rewrite Δ as follows:

$$\Delta = \underbrace{\mathrm{mmse}_n^{\mathcal{H}}(S|X^\sigma) - \mathrm{mmse}^{\mathcal{H}}(S|X^\sigma)}_{\Delta_{\mathrm{C}}} + \underbrace{\mathrm{mmse}^{\mathcal{H}}(S|X^\sigma) - \mathrm{mmse}(S|X^\sigma)}_{\Delta_{\mathrm{A}}}, \tag{40}$$

where Δ_C can be bounded using concentration inequalities, while Δ_A can be thought of as an approximation bound. Let

$$h_{\mathcal{H}}^* \in \arg\min_{h\in\mathcal{H}} \mathbb{E}[(S - h(X^\sigma))^2]. \tag{41}$$

Then

$$\Delta_{\mathrm{C}} = \inf_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n (S_i - h(X_i^\sigma))^2 - \mathbb{E}[(S - h_{\mathcal{H}}^*(X^\sigma))^2] \tag{42}$$

$$\leq \frac{1}{n}\sum_{i=1}^n (S_i - h_{\mathcal{H}}^*(X_i^\sigma))^2 - \mathbb{E}[(S - h_{\mathcal{H}}^*(X^\sigma))^2]. \tag{43}$$

Since $S \in [0,1]$ and $h_{\mathcal{H}}^* : \mathbb{R} \to [0,1]$, we have that $(S_i - h_{\mathcal{H}}^*(X_i^\sigma))^2 \leq 1$ for all $i \in \{1,\dots,n\}$. As a result, by Hoeffding's inequality [12, Sec. 4.2], with probability at least $1-\delta$,

$$\Delta_{\mathrm{C}} \leq \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{44}$$

Recall that the $\mathrm{mmse}(S|X^\sigma)$ is attained by the conditional expectation η^σ in (7) and note that

$$\begin{aligned}
\mathrm{mmse}^{\mathcal{H}}(S|X^\sigma) &= \mathbb{E}[(S - h_{\mathcal{H}}^*(X))^2] && (45)\\
&= \mathbb{E}[(S - h_{\mathcal{H}}^*(X^\sigma))^2 - (S - \eta^\sigma(X^\sigma))^2 + (S - \eta^\sigma(X^\sigma))^2] && (46)\\
&= \mathbb{E}[(\eta(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2 + 2(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))(S - \eta^\sigma(X^\sigma)) + (S - \eta^\sigma(X^\sigma))^2] && (47)\\
&= \mathbb{E}[(\eta(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2] + 2\mathbb{E}\left[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))\mathbb{E}[S - \eta^\sigma(X^\sigma)|X^\sigma]\right] + \mathbb{E}[(S - \eta^\sigma(X^\sigma))^2] && (48)\\
&= \mathbb{E}[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2] + 2\mathbb{E}\left[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))(\mathbb{E}[S|X^\sigma] - \eta^\sigma(X^\sigma))\right] + \mathbb{E}[(S - \eta(X^\sigma))^2] && (49)\\
&= \mathbb{E}[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2] + \mathbb{E}[(S - \eta^\sigma(X^\sigma))^2]. && (50)
\end{aligned}$$

Observe that (50) implies that $h_{\mathcal{H}}^* \in \arg\min_{h\in\mathcal{H}} \mathbb{E}[(\eta^\sigma(X^\sigma) - h(X^\sigma))^2]$. Thus, we have that

$$\begin{aligned}
\Delta_{\mathrm{A}} &= \mathbb{E}[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2] + \mathbb{E}[(S - \eta^\sigma(X^\sigma))^2] - \mathbb{E}[(S - \eta^\sigma(X^\sigma))^2] && (51)\\
&= \mathbb{E}[(\eta^\sigma(X^\sigma) - h_{\mathcal{H}}^*(X^\sigma))^2] && (52)\\
&= \|\eta^\sigma - h_{\mathcal{H}}^*\|_2^2, && (53)
\end{aligned}$$

where $\|\cdot\|_2$ is the 2-norm w.r.t. the distribution of X^σ, i.e.,

$$\|h\|_2^2 = \int_B |h(x)|^2 P_X(\mathrm{d}x). \tag{54}$$

Therefore, using (44) and (53), we can upper bound (40) as

$$\Delta = \Delta_{\mathrm{C}} + \Delta_{\mathrm{A}} \leq \sqrt{\frac{\log(1/\delta)}{2n}} + \|\eta^\sigma - h_{\mathcal{H}}^*\|_2^2. \tag{55}$$

B. Proof of Proposition 2

Since $\eta^\sigma(x) = s(\theta^\sigma(x))$, then there exists $\theta_L^* \in \arg\min_{\theta_L \in \mathcal{H}_L} \mathbb{E}\left[(\theta^\sigma(X^\sigma) - \theta_L(X^\sigma))^2\right]$ such that $h_{\mathcal{H}}^*(x) = s(\theta_L^*(x))$. We can then have the following bound on ϵ_A:

$$\epsilon_A = \|s \circ \theta^\sigma - s \circ \theta_L^*\|_2^2 \leq \|\theta^\sigma - \theta_L^*\|_2^2, \tag{56}$$

where the inequality follows from the fact that s is 1-Lipschitz. The function θ_L^* is of the form

$$\theta_L^*(x) = a^* x + b^*, \tag{57}$$

where

$$a^* = \text{Var}(X^\sigma)^{-1}\text{Cov}(X^\sigma, \theta(X^\sigma)), \tag{58}$$
$$b^* = \mathbb{E}[\theta(X^\sigma)] - (a^*)^T\mathbb{E}[X^\sigma]. \tag{59}$$

Then

$$\begin{aligned}\|\theta^\sigma - \theta_L^*\|_2^2 &= \mathbb{E}\left[(\theta^\sigma(X^\sigma) - \theta_L^*(X^\sigma))^2\right] \\ &= \text{Var}(\theta^\sigma(X^\sigma)) - \text{Cov}(\theta^\sigma(X^\sigma), X^\sigma)\text{Var}(X^\sigma)^{-1}\text{Cov}(X^\sigma, \theta^\sigma(X^\sigma)).\end{aligned} \tag{60}$$

C. Proof of Proposition 3

Consider setting (ii) first, where $(X|S = s) \sim N(\mu_s, \Sigma)$ for mean $\mu_s \in \mathbb{R}^d$ and a symmetric, positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Let $\tilde{\Sigma} \coloneqq \Sigma + \sigma^2 I$. We have that for $i \in \{0, 1\}$,

$$f_i^\sigma(x) = \frac{1}{(2\pi)^{d/2}|\tilde{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\tilde{\Sigma}^{-1}(x-\mu_i)}. \tag{61}$$

By (8), for all $x \in \mathbb{R}^d$,

$$\theta^\sigma(x) = (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}x + \frac{1}{2}\left(\mu_0^{\mathrm{T}}\tilde{\Sigma}^{-1}\mu_0 - \mu_1^{\mathrm{T}}\tilde{\Sigma}^{-1}\mu_1\right) + \log\left(\frac{p}{\bar{p}}\right). \tag{62}$$

Computing each term in (17), we get

$$\text{Var}(\theta^\sigma(X^\sigma)) = (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}(\text{Var}(X) + \sigma^2 I)\tilde{\Sigma}^{-1}(\mu_1 - \mu_0), \tag{63}$$

$$\text{Cov}(\theta^\sigma(X^\sigma), X^\sigma) = \text{Cov}(X^\sigma, \theta^\sigma(X^\sigma))^{\mathrm{T}} = (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}(\text{Var}(X) + \sigma^2 I), \tag{64}$$

$$\text{Var}(X^\sigma) = \text{Var}(X) + \sigma^2 I, \tag{65}$$

which yields

$$\begin{aligned}\epsilon_A \leq\ & (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}(\text{Var}(X) + \sigma^2 I)\tilde{\Sigma}^{-1}(\mu_1 - \mu_0) \\ & - (\mu_1 - \mu_0)^{\mathrm{T}}\tilde{\Sigma}^{-1}(\text{Var}(X) + \sigma^2 I)(\text{Var}(X) + \sigma^2 I)^{-1}(\text{Var}(X) + \sigma^2 I)\tilde{\Sigma}^{-1}(\mu_1 - \mu_0) = 0.\end{aligned} \tag{66}$$

This implies $\epsilon_A = 0$.

Note that once noise is added to X, setting (i), where $X = aS + b$ for $a, b \in \mathbb{R}^d$, becomes a special case of setting (ii) with $\mu_1 = a + b$, $\mu_0 = b$, and $\tilde{\Sigma} = \sigma^2 I$. Therefore, the same result follows.

D. Proof of Theorem 1

We have that

$$f_1^\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\left[p_N e^{-x^2/2\sigma^2} + \overline{p_N} e^{-(x-1)^2/2\sigma^2}\right] \tag{67}$$

and

$$f_0^\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\left[\overline{p_N} e^{-x^2/2\sigma^2} + p_N e^{-(x-1)^2/2\sigma^2}\right] \tag{68}$$

By (8), for all $x \in \mathbb{R}$,

$$\theta^\sigma(x) = \log\left(\frac{p\left[p_N e^{-x^2/2\sigma^2} + \overline{p_N} e^{-(x-1)^2/2\sigma^2}\right]}{\overline{p}\left[\overline{p_N} e^{-x^2/2\sigma^2} + p_N e^{-(x-1)^2/2\sigma^2}\right]}\right) = \log\left(\frac{p\left[p_N + \overline{p_N} e^{(2x-1)/2\sigma^2}\right]}{\overline{p}\left[\overline{p_N} + p_N e^{(2x-1)/2\sigma^2}\right]}\right). \tag{69}$$

Note that

$$P_X(x) = \begin{cases} q \coloneqq p_N\overline{p} + \overline{p_N}p, & x = 1 \\ \overline{q} \coloneqq 1 - q, & x = 0. \end{cases} \tag{70}$$

Let

$$I_0 \coloneqq \mathbb{E}[\theta^\sigma(X^\sigma)|X = 0] = \int_{-\infty}^{\infty} \theta^\sigma(\sigma z) f_Z(z) dz, \tag{71}$$

$$I_1 \coloneqq \mathbb{E}[\theta^\sigma(X^\sigma)|X = 1] = \int_{-\infty}^{\infty} \theta^\sigma(1 + \sigma z) f_Z(z) dz. \tag{72}$$

Then

$$\text{Var}(\mathbb{E}\left[\theta^\sigma(X^\sigma)|X\right]) = q\overline{q}(I_1 - I_0)^2. \tag{73}$$

Rewriting (69) as

$$\theta^\sigma(x) = \log\left(\frac{p}{\overline{p}}\right) + \log\left(1 + \overline{p_N}(\beta_x - 1)\right) - \log\left(1 + p_N(\beta_x - 1)\right), \tag{74}$$

where $\beta_x = e^{(2x-1)/2\sigma^2}$, and using the Maclaurin series expansion of $\log(1+x)$ yields

$$\theta^\sigma(x) = \log\left(\frac{p}{\overline{p}}\right) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n} (\beta_x - 1)^n [p_N^n - (1 - p_N)^n]. \tag{75}$$

Then we can write $\tilde{I_0} \coloneqq I_0 - \log(p/\overline{p})$ as follows:

$$\begin{aligned}
\tilde{I_0} &= \int_{-\infty}^{\infty} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \left(e^{(2\sigma z - 1)/2\sigma^2} - 1\right)^n [p_N^n - (1 - p_N)^n] \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz \\
&= \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \int_{-\infty}^{\infty} \left(e^{(2\sigma z - 1)/2\sigma^2} - 1\right)^n \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz \\
&\overset{(a)}{=} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \int_{-\infty}^{\infty} \sum_{k=0}^{n} \binom{n}{k} e^{(n-k)(2\sigma z - 1)/2\sigma^2} (-1)^k \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz \\
&= \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \int_{-\infty}^{\infty} \exp\left(\frac{-(z\sigma - (n-k))^2 + (n-k)^2 - (n-k)}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}} \, dz \\
&= \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \sum_{k=0}^{n} \binom{n}{k} (-1)^k e^{(k-n)(1+k-n)/2\sigma^2} \\
&\overset{(b)}{=} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \sum_{m=0}^{\infty} \frac{(k-n)^m (1+k-n)^m}{m! 2^m \sigma^{2m}} \\
&= \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{(k-n)^m (1+k-n)^m}{m! 2^m \sigma^{2m}},
\end{aligned}$$

where (a) follows from the Binomial theorem and (b) follows from the Maclaurin series expansion of e^x. Considering the first few values of m, we obtain

$$m = 0: \quad \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \sum_{k=0}^{n} \binom{n}{k} (-1)^k = 0 \tag{76}$$

$$m = 1: \quad \sum_{n=1}^{\infty} \frac{(-1)^n}{n} [p_N^n - (1 - p_N)^n] \underbrace{\sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{(k-n)(1+k-n)}{2\sigma^2}}_{= \begin{cases} \frac{1}{\sigma^2} & n = 2, \\ 0 & \text{o.w.} \end{cases}} = \frac{2p_N - 1}{2\sigma^2} \tag{77}$$

$$m=2: \quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\underbrace{\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(k-n)^2(1+k-n)^2}{8\sigma^4}}_{=\begin{cases}\frac{1}{2\sigma^4} & n=2,\\ \frac{3}{\sigma^4} & n=3,4,\\ 0 & \text{o.w.}\end{cases}}=\frac{p_N(1-3p_N+2p_N^2)}{2\sigma^4}. \tag{78}$$

Therefore,

$$I_0\approx\log\left(\frac{p}{\overline{p}}\right)+\frac{2p_N-1}{2\sigma^2}+\frac{p_N(1-3p_N+2p_N^2)}{2\sigma^4}. \tag{79}$$

Similarly, we can write I_1 as

$$I_1=\log\left(\frac{p}{\overline{p}}\right)+\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(n-k)^m(1+n-k)^m}{m!2^m\sigma^{2m}}. \tag{80}$$

Again considering only the first few values of m, we get

$$m=0: \quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\sum_{k=0}^{n}\binom{n}{k}(-1)^k=0 \tag{81}$$

$$m=1: \quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\underbrace{\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(n-k)(1+n-k)}{2\sigma^2}}_{=\begin{cases}\frac{1}{\sigma^2} & n=1,2,\\ 0 & \text{o.w.}\end{cases}}=\frac{1-2p_N}{2\sigma^2} \tag{82}$$

$$m=2: \quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\underbrace{\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(n-k)^2(1+n-k)^2}{8\sigma^4}}_{=\begin{cases}\frac{1}{2\sigma^4} & n=1,\\ \frac{7}{2\sigma^4} & n=2,\\ \frac{6}{\sigma^4} & n=3,\\ \frac{3}{\sigma^4} & n=4,\\ 0 & \text{o.w.}\end{cases}}=\frac{-p_N(1-3p_N+2p_N^2)}{2\sigma^4}. \tag{83}$$

Therefore,

$$I_1\approx\log\left(\frac{p}{\overline{p}}\right)+\frac{1-2p_N}{2\sigma^2}-\frac{p_N(1-3p_N+2p_N^2)}{2\sigma^4}. \tag{84}$$

Hence,

$$\begin{aligned}\text{Var}(\mathbb{E}\left[\theta^\sigma(X^\sigma)|X\right]) &= q\overline{q}(I_1-I_0)^2\\ &\approx q\overline{q}\left(\frac{1-2p_N}{\sigma^2}-\frac{p_N(1-3p_N+2p_N^2)}{\sigma^4}\right)^2\\ &= q\overline{q}\left(\frac{(1-2p_N)^2}{\sigma^4}-\frac{2p_N(1-2p_N)(1-3p_N+2p_N^2)}{\sigma^6}+\frac{p_N^2(1-3p_N+2p_N^2)^2}{\sigma^8}\right).\end{aligned} \tag{85}$$

Let

$$I_{02}\coloneqq\mathbb{E}[\theta^\sigma(X^\sigma)^2|X=0]=\int_{-\infty}^{\infty}\theta^\sigma(\sigma z)^2f_Z(z)dz, \tag{86}$$

$$I_{12}\coloneqq\mathbb{E}[\theta^\sigma(X^\sigma)^2|X=1]=\int_{-\infty}^{\infty}\theta^\sigma(1+\sigma z)^2f_Z(z)dz. \tag{87}$$

Then we also have that

$$\mathbb{E}\left[\text{Var}(\theta^{\sigma}(X^{\sigma})|X)\right] = \overline{q}I_{02} + qI_{12} - \overline{q}I_0^2 - qI_1^2. \tag{88}$$

Again using the series expansion of θ^{σ} in (75), we can write I_{02} as

$$\begin{aligned}
I_{02} &= \log^2\left(\frac{p}{\overline{p}}\right) + 2\log\left(\frac{p}{\overline{p}}\right)\left(I_0 - \log\left(\frac{p}{\overline{p}}\right)\right) \\
&\quad + \int_{-\infty}^{\infty}\left(\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\left(e^{(2\sigma z-1)/2\sigma^2}-1\right)^n\left[p_N^n-(1-p_N)^n\right]\right)^2\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \\
&= \log^2\left(\frac{p}{\overline{p}}\right) + 2\log\left(\frac{p}{\overline{p}}\right)\left(I_0 - \log\left(\frac{p}{\overline{p}}\right)\right) \\
&\quad + \sum_{n=1}^{\infty}\frac{[p_N^n-(1-p_N)^n]^2}{n^2}\int_{-\infty}^{\infty}\left(e^{(2\sigma z-1)/2\sigma^2}-1\right)^{2n}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \\
&\quad + 2\sum_{j<n}\frac{(-1)^{n+j}[p_N^n-(1-p_N)^n][p_N^j-(1-p_N)^j]}{nj}\int_{-\infty}^{\infty}\left(e^{(2\sigma z-1)/2\sigma^2}-1\right)^{n+j}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \\
&\overset{(a)}{=} \log^2\left(\frac{p}{\overline{p}}\right) + 2\log\left(\frac{p}{\overline{p}}\right)\left(I_0 - \log\left(\frac{p}{\overline{p}}\right)\right) \\
&\quad + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\left[\frac{[p_N^n-(1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)^m(1+k-2n)^m}{m!2^m\sigma^{2m}}\right. \\
&\quad \left. + 2\sum_{j<n}\frac{(-1)^{n+j}[p_N^n-(1-p_N)^n][p_N^j-(1-p_N)^j]}{nj}\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)^m(1+k-n-j)^m}{m!2^m\sigma^{2m}}\right],
\end{aligned} \tag{89}$$

where (a) follows from similar analysis to that used for I_0. Considering just the second term in (89) and analyzing it for the first few values of m yields

$$m=0: \quad \sum_{n=1}^{\infty}\frac{[p_N^n-(1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k = 0 \tag{90}$$

$$m=1: \quad \sum_{n=1}^{\infty}\frac{[p_N^n-(1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)(1+k-2n)}{2\sigma^2} = \frac{(2p_N-1)^2}{2\sigma^2} \tag{91}$$

$$m=2: \quad \sum_{n=1}^{\infty}\frac{[p_N^n-(1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)^2(1+k-2n)^2}{8\sigma^4} = \frac{5(1-2p_N)^2}{4\sigma^4}. \tag{92}$$

Doing the same for the third term in (89) yields

$$m=0: \quad 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n-(1-p_N)^n][p_N^j-(1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k = 0 \tag{93}$$

$$\begin{aligned}
m=1: \quad & 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n-(1-p_N)^n][p_N^j-(1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)(1+k-n-j)}{2\sigma^2} \\
&= \frac{2(2p_N-1)^2}{\sigma^2}
\end{aligned} \tag{94}$$

$$\begin{aligned}
m=2: \quad & 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n-(1-p_N)^n][p_N^j-(1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)^2(1+k-n-j)^2}{8\sigma^4} \\
&= \frac{(1-2p_N)^2(3-4p_N+4p_N^2)}{2\sigma^4}.
\end{aligned} \tag{95}$$

Therefore,

$$\begin{aligned}
I_{02} &\approx \log^2\left(\frac{p}{\overline{p}}\right) + 2\log\left(\frac{p}{\overline{p}}\right)\left(\frac{2p_N-1}{2\sigma^2} + \frac{p_N(1-3p_N+2p_N^2)}{2\sigma^4}\right) \\
&\quad + \frac{5(2p_N-1)^2}{2\sigma^2} + \frac{(1-2p_N)^2(11-8p_N+8p_N^2)}{4\sigma^4}.
\end{aligned} \tag{96}$$

Similarly, we can write I_{12} as

$$\begin{aligned}I_{12} &= \log^2\left(\frac{p}{\overline{p}}\right) + 2\log\left(\frac{p}{\overline{p}}\right)\left(I_1 - \log\left(\frac{p}{\overline{p}}\right)\right) \\ &+ \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\Bigg[\frac{[p_N^n - (1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)^m(k-2n-1)^m}{m!2^m\sigma^{2m}} \\ &+ 2\sum_{j<n}\frac{(-1)^{n+j}[p_N^n - (1-p_N)^n][p_N^j - (1-p_N)^j]}{nj}\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)^m(k-n-j-1)^m}{m!2^m\sigma^{2m}}\Bigg], \end{aligned} \tag{97}$$

Considering just the second term in (97) and analyzing it for the first few values of m yields

$$m=0: \quad \sum_{n=1}^{\infty}\frac{[p_N^n - (1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k = 0 \tag{98}$$

$$m=1: \quad \sum_{n=1}^{\infty}\frac{[p_N^n - (1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)(k-2n-1)}{2\sigma^2} = \frac{(2p_N-1)^2}{\sigma^2} \tag{99}$$

$$m=2: \quad \sum_{n=1}^{\infty}\frac{[p_N^n - (1-p_N)^n]^2}{n^2}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\frac{(k-2n)^2(k-2n-1)^2}{8\sigma^4} = \frac{17(1-2p_N)^2}{4\sigma^4}. \tag{100}$$

Doing the same for the third term in (97) yields

$$m=0: \quad 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n - (1-p_N)^n][p_N^j - (1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k = 0 \tag{101}$$

$$\begin{aligned}m=1: \quad & 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n - (1-p_N)^n][p_N^j - (1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)(k-n-j-1)}{2\sigma^2} \\ &= \frac{2(2p_N-1)^2}{\sigma^2}\end{aligned} \tag{102}$$

$$\begin{aligned}m=2: \quad & 2\sum_{n=1}^{\infty}\sum_{j<n}\frac{(-1)^{n+j}}{nj}[p_N^n - (1-p_N)^n][p_N^j - (1-p_N)^j]\sum_{k=0}^{n+j}\binom{n+j}{k}(-1)^k\frac{(k-n-j)^2(k-n-j-1)^2}{8\sigma^4} \\ &= \frac{(1-2p_N)^2(9-4p_N+4p_N^2)}{2\sigma^4}.\end{aligned} \tag{103}$$

Therefore,

$$\begin{aligned}I_{12} \approx \log^2\left(\frac{p}{\overline{p}}\right) &+ 2\log\left(\frac{p}{\overline{p}}\right)\left(\frac{1-2p_N}{2\sigma^2} - \frac{p_N(1-3p_N+2p_N^2)}{2\sigma^4}\right) \\ &+ \frac{3(2p_N-1)^2}{\sigma^2} + \frac{(1-2p_N)^2(35-8p_N+8p_N^2)}{4\sigma^4}.\end{aligned} \tag{104}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}(\theta^\sigma(X^\sigma)|X)\right] &= \overline{q}(I_{02} - I_0^2) + q(I_{12} - I_1^2) \\ &\approx \overline{q}\left(\frac{5(2p_N-1)^2}{2\sigma^2} + \frac{(1-2p_N)^2(10-8p_N+8p_N^2)}{4\sigma^4} - \frac{2p_N(2p_N-1)(1-3p_N+2p_N^2)}{4\sigma^6}\right. \\ &\qquad \left. - \frac{p_N^2(1-3p_N+2p_N^2)^2}{4\sigma^8}\right) \\ &\quad + q\left(\frac{3(2p_N-1)^2}{\sigma^2} + \frac{(1-2p_N)^2(34-8p_N+8p_N^2)}{4\sigma^4} + \frac{2p_N(1-2p_N)(1-3p_N+2p_N^2)}{4\sigma^6}\right. \\ &\qquad \left. - \frac{p_N^2(1-3p_N+2p_N^2)^2}{4\sigma^8}\right) \\ &= \frac{5(2p_N-1)^2}{2\sigma^2} + \frac{(1-2p_N)^2(10-8p_N+8p_N^2)}{4\sigma^4} - \frac{2p_N(2p_N-1)(1-3p_N+2p_N^2)}{4\sigma^6} \\ &\qquad - \frac{p_N^2(1-3p_N+2p_N^2)^2}{4\sigma^8} + q(2p_N-1)^2\left(\frac{1}{2\sigma^2} + \frac{8}{\sigma^4}\right).\end{aligned} \tag{105}$$

Finally, we have that

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma) &= \mathrm{Cov}(\theta^\sigma(X^\sigma), X) + \sigma \mathrm{Cov}(\theta^\sigma(X^\sigma), Z) && (106)\\ &= q\overline{q}(I_1 - I_0) + \sigma\left(\overline{q}\mathbb{E}[Z\theta^\sigma(\sigma Z)] + q\mathbb{E}[Z\theta^\sigma(1+\sigma Z)]\right). && (107)\end{aligned}$$

Observe that

$$\begin{aligned}E[Z\theta^\sigma(\sigma Z)] &= \int_{-\infty}^{\infty} z\theta^\sigma(\sigma z) f_Z(z) dz && (108)\\ &= \int_0^{\infty} z\theta^\sigma(\sigma z) f_Z(z) dz - \int_0^{\infty} z\theta^\sigma(-\sigma z) f_Z(-z) dz && (109)\\ &= \int_0^{\infty} z(\theta^\sigma(\sigma z) - \theta^\sigma(-\sigma z)) f_Z(z) dz && (110)\\ &= \int_0^{\infty} z\left(\log\left(\frac{\left[p_N + \overline{p_N} e^{(2\sigma z-1)/2\sigma^2}\right]\left[\overline{p_N} + p_N e^{(-2\sigma z-1)/2\sigma^2}\right]}{\left[p_N + \overline{p_N} e^{(-2\sigma z-1)/2\sigma^2}\right]\left[\overline{p_N} + p_N e^{(2\sigma z-1)/2\sigma^2}\right]}\right)\right) f_Z(z) dz && (111)\\ &= \int_0^{\infty} z\left(\log\left(\frac{p_N\overline{p_N}(1+e^{-1/\sigma^2}) + p_N^2 e^{(-2\sigma z-1)/2\sigma^2} + \overline{p_N}^2 e^{(2\sigma z-1)/2\sigma^2}}{p_N\overline{p_N}(1+e^{-1/\sigma^2}) + p_N^2 e^{(2\sigma z-1)/2\sigma^2} + \overline{p_N}^2 e^{(-2\sigma z-1)/2\sigma^2}}\right)\right) f_Z(z) dz && (112)\\ &= \int_0^{\infty} z\left(\log\left(\frac{p_N\overline{p_N}(1+e^{1/\sigma^2}) + p_N^2 e^{(-2\sigma z+1)/2\sigma^2} + \overline{p_N}^2 e^{(2\sigma z+1)/2\sigma^2}}{p_N\overline{p_N}(1+e^{1/\sigma^2}) + p_N^2 e^{(2\sigma z+1)/2\sigma^2} + \overline{p_N}^2 e^{(-2\sigma z+1)/2\sigma^2}}\right)\right) f_Z(z) dz && (113)\\ &= \int_0^{\infty} z\left(\log\left(\frac{\left[p_N + \overline{p_N} e^{(2\sigma z+1)/2\sigma^2}\right]\left[\overline{p_N} + p_N e^{(-2\sigma z+1)/2\sigma^2}\right]}{\left[p_N + \overline{p_N} e^{(-2\sigma z+1)/2\sigma^2}\right]\left[\overline{p_N} + p_N e^{(2\sigma z+1)/2\sigma^2}\right]}\right)\right) f_Z(z) dz && (114)\\ &= \int_0^{\infty} z(\theta^\sigma(\sigma z+1) - \theta^\sigma(-\sigma z+1)) f_Z(z) dz && (115)\\ &= \int_{-\infty}^{\infty} z\theta^\sigma(\sigma z+1) f_Z(z) dz && (116)\\ &= E[Z\theta^\sigma(\sigma Z+1)]. && (117)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma) &= q\overline{q}(I_1 - I_0) + \sigma\left(\overline{q}\mathbb{E}[Z\theta^\sigma(\sigma Z)] + q\mathbb{E}[Z\theta^\sigma(1+\sigma Z)]\right) && (118)\\ &= q\overline{q}(I_1 - I_0) + \sigma\mathbb{E}[Z\theta^\sigma(\sigma Z)]. && (119)\end{aligned}$$

Using similar analysis to that above and the series expansion in (75), we can write $\mathbb{E}[Z\theta^\sigma(\sigma Z)]$ as

$$\begin{aligned}\mathbb{E}[Z\theta^\sigma(\sigma Z)] &= \int_{-\infty}^{\infty} z \sum_{n=1}^{\infty} \frac{(-1)^n}{n}\left(e^{(2\sigma z-1)/2\sigma^2} - 1\right)^n \left[p_N^n - (1-p_N)^n\right] \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz\\ &= \sum_{n=1}^{\infty} \frac{(-1)^n}{n}\left[p_N^n - (1-p_N)^n\right] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \int_{-\infty}^{\infty} \exp\left(\frac{-(z\sigma - (n-k))^2 + (n-k)^2 - (n-k)}{2\sigma^2}\right) \frac{z}{\sqrt{2\pi}}\, dz\\ &= \sum_{n=1}^{\infty} \frac{(-1)^n}{n}\left[p_N^n - (1-p_N)^n\right] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{n-k}{\sigma} e^{(k-n)(1+k-n)/2\sigma^2}\\ &= \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \frac{(-1)^n}{n}\left[p_N^n - (1-p_N)^n\right] \sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{n-k}{\sigma} \frac{(k-n)^m (1+k-n)^m}{m! 2^m \sigma^{2m}}.\end{aligned}$$

Considering only the first few values of m, we get

$$m = 0: \quad \sum_{n=1}^{\infty} \frac{(-1)^n}{n}\left[p_N^n - (1-p_N)^n\right] \underbrace{\sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{(n-k)}{\sigma}}_{= \begin{cases} \frac{1}{\sigma} & n=1, \\ 0 & \text{o.w.} \end{cases}} = \frac{1-2p_N}{\sigma} \tag{120}$$

$$m=1:\quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\underbrace{\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(n-k)}{\sigma}\frac{(k-n)(1+k-n)}{2\sigma^2}}_{=\begin{cases}\frac{2}{\sigma^3} & n=2,\\ \frac{3}{\sigma^3} & n=3,\\ 0 & \text{o.w.}\end{cases}}=\frac{p_N(-1+3p_N-2p_N^2)}{\sigma^3} \tag{121}$$

$$m=2:\quad \sum_{n=1}^{\infty}\frac{(-1)^n}{n}[p_N^n-(1-p_N)^n]\underbrace{\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{(n-k)}{\sigma^2}\frac{(k-n)^2(1+k-n)^2}{8\sigma^4}}_{=\begin{cases}\frac{1}{\sigma^5} & n=2,\\ \frac{21}{2\sigma^5} & n=3,\\ \frac{24}{\sigma^5} & n=4,\\ \frac{15}{\sigma^5} & n=5,\\ 0 & \text{o.w.}\end{cases}}=\frac{-p_N(1-9p_N+26p_N^2-30p_N^3+12p_N^4)}{2\sigma^5}. \tag{122}$$

Therefore,

$$\mathbb{E}[Z\theta^\sigma(\sigma Z)]\approx\frac{1-2p_N}{\sigma}+\frac{p_N(-1+3p_N-2p_N^2)}{\sigma^3}-\frac{p_N(1-9p_N+26p_N^2-30p_N^3+12p_N^4)}{2\sigma^5}. \tag{123}$$

Hence,

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma),X^\sigma)&=q\overline{q}(I_1-I_0)+\sigma\mathbb{E}[Z\theta^\sigma(\sigma Z)]\\&\approx q\overline{q}\left(\frac{1-2p_N}{\sigma^2}-\frac{p_N(1-3p_N+2p_N^2)}{\sigma^4}\right)+1-2p_N+\frac{p_N(-1+3p_N-2p_N^2)}{\sigma^2}\\&\quad-\frac{p_N(1-9p_N+26p_N^2-30p_N^3+12p_N^4)}{2\sigma^4}\\&=1-2p_N+\frac{p_N(-1+3p_N-2p_N^2)+q\overline{q}(1-2p_N)}{\sigma^2}\\&\quad-\frac{p_N(1-9p_N+26p_N^2-30p_N^3+12p_N^4)+2q\overline{q}p_N(1-3p_N+2p_N^2)}{2\sigma^4}.\end{aligned} \tag{124}$$

Note that

$$\mathrm{Var}(X^\sigma)=\mathrm{Var}(X)+\sigma^2=q\overline{q}+\sigma^2. \tag{125}$$

Putting together (85), (105), (124) and (125), we obtain

$$\epsilon_A\le\frac{2(1-2p_N)^2}{\sigma^2+p_N(1-2p)^2\overline{p_N}+p\overline{p}}-\frac{(1-2p_N)^2[4p_Np^2-(6+4p_N)p-9+p_N]}{2\sigma^2[p_N(1-2p)^2\overline{p_N}+p\overline{p}]+2\sigma^4}+\mathcal{O}\left(\frac{C(p,p_N)(1-2p_N)^2}{2\sigma^4[p_N(1-2p)^2\overline{p_N}+p\overline{p}]+2\sigma^6}\right) \tag{126}$$

for some constant $C(p,p_N)$.

E. Proof of Theorem 2

For $s\in\{0,1\}$, let $\tilde{\Sigma}_s\coloneqq\Sigma_s+\sigma^2I$. Then

$$f_s^\sigma(x)=\frac{1}{(2\pi)^{d/2}|\tilde{\Sigma}_s|^{1/2}}e^{-\frac{1}{2}(x-\mu_s)^{\mathrm{T}}\tilde{\Sigma}_s^{-1}(x-\mu_s)}. \tag{127}$$

By (8), for all $x\in\mathbb{R}^d$,

$$\theta^\sigma(x)=x^{\mathrm{T}}Ax+b^{\mathrm{T}}x+c, \tag{128}$$

where

$$A = \frac{1}{2}\left(\tilde{\Sigma}_0^{-1} - \tilde{\Sigma}_1^{-1}\right),$$
$$b = \tilde{\Sigma}_1^{-1}\mu_1 - \tilde{\Sigma}_0^{-1}\mu_0,$$
$$c = \frac{1}{2}\mu_0^{\mathrm{T}}\tilde{\Sigma}_0^{-1}\mu_0 - \frac{1}{2}\mu_1^{\mathrm{T}}\tilde{\Sigma}_1^{-1}\mu_1 + \frac{1}{2}\log\left(\frac{|\tilde{\Sigma}_0|}{|\tilde{\Sigma}_1|}\right) + \log\left(\frac{p}{1-p}\right).$$

By the law of total variance,

$$\mathrm{Var}(\theta^\sigma(X^\sigma)) = \mathbb{E}[\mathrm{Var}(\theta^\sigma(X^\sigma)|S)] + \mathrm{Var}(\mathbb{E}[\theta^\sigma(X^\sigma)|S]). \tag{129}$$

From [13, Thm. 3.2b.3], we have that for $s \in \{0,1\}$,

$$M_s \coloneqq \mathbb{E}[\theta^\sigma(X^\sigma)|S=s] = \sum_{j=1}^{d}(\lambda_j^{(s)}) + (c + b^{\mathrm{T}}\mu_s + \mu_s^{\mathrm{T}}A\mu_s), \tag{130}$$

$$V_s \coloneqq \mathrm{Var}(\theta^\sigma(X^\sigma)|S=s) = \sum_{j=1}^{d} 2(\lambda_j^{(s)})^2 + (u_j^{(s)})^2, \tag{131}$$

where $\lambda_j^{(s)}$, $j \in \{1,\dots,d\}$, are the eigenvalues of $\tilde{\Sigma}_s^{1/2}A\tilde{\Sigma}_s^{1/2}$ with corresponding eigenvectors as the columns of a matrix Q_s, i.e., $Q_s^{\mathrm{T}}\tilde{\Sigma}_s^{1/2}A\tilde{\Sigma}_s^{1/2}Q_s = \mathrm{diag}(\lambda_1^{(s)},\dots,\lambda_d^{(s)})$, and

$$u^{(s)} = (u_1^{(s)},\dots,u_d^{(s)})^{\mathrm{T}} = Q_s^{\mathrm{T}}(\tilde{\Sigma}_s^{1/2}b + 2\tilde{\Sigma}_s^{1/2}A\mu_s). \tag{132}$$

We can therefore compute the first term in (129) as

$$\mathbb{E}[\mathrm{Var}(\theta^\sigma(X^\sigma)|S)] = \mathbb{E}\left[SV_1 + (1-S)V_0\right] = pV_1 + (1-p)V_0 \tag{133}$$

and the second term as

$$\mathrm{Var}(\mathbb{E}[\theta^\sigma(X^\sigma)|S]) = \mathrm{Var}(S)\left(M_1 - M_0\right)^2 = p(1-p)\left(M_1 - M_0\right)^2. \tag{134}$$

Therefore,

$$\mathrm{Var}(\theta^\sigma(X^\sigma)) = pV_1 + (1-p)V_0 + p(1-p)\left(M_1 - M_0\right)^2. \tag{135}$$

Next, we compute $\mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma)$ as

$$\mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma) = \mathrm{Cov}(X^{\mathrm{T}}AX + 2\sigma Z^{\mathrm{T}}AX + \sigma^2 Z^{\mathrm{T}}AZ + b^{\mathrm{T}}X + \sigma b^{\mathrm{T}}Z, X + \sigma Z) \tag{136}$$
$$= \mathrm{Cov}(X^{\mathrm{T}}AX, X) + b^{\mathrm{T}}\mathrm{Var}(X) + 2\sigma^2\mathrm{Cov}(Z^{\mathrm{T}}AX, Z) + \sigma^2 b^{\mathrm{T}}\mathrm{Var}(Z) \tag{137}$$
$$= \mathrm{Cov}(X^{\mathrm{T}}AX, X) + b^{\mathrm{T}}\mathrm{Var}(X) + 2\sigma^2\mathrm{Cov}(Z^{\mathrm{T}}AX, Z) + \sigma^2 b^{\mathrm{T}}. \tag{138}$$

Note that

$$\mathrm{Cov}(Z^{\mathrm{T}}AX, Z) = \mathbb{E}[Z^{\mathrm{T}}AXZ^{\mathrm{T}}] = \mathbb{E}[X^{\mathrm{T}}AZZ^{\mathrm{T}}] = \mathbb{E}[X]^{\mathrm{T}}A\mathbb{E}[ZZ^{\mathrm{T}}] = [p\mu_1 + (1-p)\mu_0]^{\mathrm{T}}A \tag{139}$$

and

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|S)] + \mathrm{Var}(\mathbb{E}[X|S]) \tag{140}$$
$$= \mathbb{E}[S\Sigma_1 + (1-S)\Sigma_0] + \mathrm{Var}\left(S\mu_1 + (1-S)\mu_0\right) \tag{141}$$
$$= p\Sigma_1 + (1-p)\Sigma_0 + p(1-p)(\mu_1 - \mu_0)(\mu_1 - \mu_0)^{\mathrm{T}}. \tag{142}$$

By the law of total covariance, we also have that

$$\mathrm{Cov}(X^{\mathrm{T}}AX, X) = \mathbb{E}[\mathrm{Cov}(X^{\mathrm{T}}AX, X|S)] + \mathrm{Cov}(\mathbb{E}[X^{\mathrm{T}}AX|S], \mathbb{E}[X|S]). \tag{143}$$

Focusing on the first term, let $\tilde{X}_s \coloneqq (X|S=s) - \mu_s$ for $s \in \{0,1\}$. Then $\tilde{X}_s \sim \mathcal{N}(0,\Sigma_s)$, and hence

$$\mathrm{Cov}(X^{\mathrm{T}}AX, X|S=s) = \mathrm{Cov}\left((\tilde{X}_s + \mu_s)^{\mathrm{T}}A(\tilde{X}_s + \mu_s), \tilde{X}_s + \mu_s\right) \tag{144}$$
$$= \mathrm{Cov}\left(\tilde{X}_s^{\mathrm{T}}A\tilde{X}_s, \tilde{X}_s\right) + 2\mu_s^{\mathrm{T}}A\mathrm{Var}\left(\tilde{X}_s\right) \tag{145}$$
$$= 2\mu_s^{\mathrm{T}}A\Sigma_s. \tag{146}$$

Therefore,

$$\mathbb{E}[\mathrm{Cov}(X^{\mathrm{T}}AX, X|S)] = 2p\mu_1^{\mathrm{T}}A\Sigma_1 + 2(1-p)\mu_0^{\mathrm{T}}A\Sigma_0. \tag{147}$$

Now, focusing on the second term in (143), from [13, Thm. 3.2b.2], we have that

$$\mathbb{E}[X^{\mathrm{T}}AX|S=s] = \mathrm{tr}(A\Sigma_s) + \mu_s^{\mathrm{T}}A\mu_s. \tag{148}$$

Therefore,

$$\mathbb{E}[X^{\mathrm{T}}AX|S] = S\left(\mathrm{tr}(A(\Sigma_1-\Sigma_0)) + \mu_1^{\mathrm{T}}A\mu_1 - \mu_0^{\mathrm{T}}A\mu_0\right) + \mathrm{tr}(A\Sigma_0) + \mu_0^{\mathrm{T}}A\mu_0, \tag{149}$$

and hence

$$\begin{aligned}\mathrm{Cov}(\mathbb{E}[X^{\mathrm{T}}AX|S], \mathbb{E}[X|S]) &= \mathrm{Cov}\left(S\left(\mathrm{tr}(A(\Sigma_1-\Sigma_0)) + \mu_1^{\mathrm{T}}A\mu_1 - \mu_0^{\mathrm{T}}A\mu_0\right), S(\mu_1-\mu_0)\right) &(150)\\ &= \left(\mathrm{tr}(A(\Sigma_1-\Sigma_0)) + \mu_1^{\mathrm{T}}A\mu_1 - \mu_0^{\mathrm{T}}A\mu_0\right)p(1-p)(\mu_1\mu_0)^{\mathrm{T}}. &(151)\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma) = {}& 2p\mu_1^{\mathrm{T}}A\Sigma_1 + 2(1-p)\mu_0^{\mathrm{T}}A\Sigma_0\\ &+ \left(\mathrm{tr}(A(\Sigma_1-\Sigma_0)) + \mu_1^{\mathrm{T}}A\mu_1 - \mu_0^{\mathrm{T}}A\mu_0\right)p(1-p)(\mu_1-\mu_0)^{\mathrm{T}}\\ &+ b^{\mathrm{T}}\left[p\Sigma_1 + (1-p)\Sigma_0 + p(1-p)(\mu_1-\mu_0)(\mu_1-\mu_0)^{\mathrm{T}}\right]\\ &+ 2\sigma^2\left[p\mu_1 + (1-p)\mu_0\right]^{\mathrm{T}}A + \sigma^2 b^{\mathrm{T}}. \end{aligned}\tag{152}$$

Note that $\mathrm{Cov}(X^\sigma, \theta^\sigma(X^\sigma)) = \mathrm{Cov}(\theta^\sigma(X^\sigma), X^\sigma)^{\mathrm{T}}$. Finally, we have that

$$\mathrm{Var}(X^\sigma) = \mathrm{Var}(X) + \sigma^2\mathrm{Var}(Z) = p\Sigma_1 + (1-p)\Sigma_0 + p(1-p)(\mu_1-\mu_0)(\mu_1-\mu_0)^{\mathrm{T}} + \sigma^2 I. \tag{153}$$

Combining (135), (152) and (153), we obtain the result.

F. Proof of Corollary 1

When $\Sigma_s = \sigma_s^2 I$ for $s \in \{0,1\}$, (128) simplifies to

$$\theta^\sigma(x) = x^{\mathrm{T}}Ax + b^{\mathrm{T}}x + c, \tag{154}$$

where

$$\begin{aligned}A &= \frac{1}{2}\left((\sigma_0^2+\sigma^2)^{-1} - (\sigma_1^2+\sigma^2)^{-1}\right)I = \frac{\sigma_1^2-\sigma_0^2}{2(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}I =: aI,\\ b &= (\sigma_1^2+\sigma^2)^{-1}\mu_1 - (\sigma_0^2+\sigma^2)^{-1}\mu_0 = \frac{(\sigma_0^2+\sigma^2)\mu_1 - (\sigma_1^2+\sigma^2)\mu_0}{(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)},\\ c &= \frac{1}{2}(\sigma_0^2+\sigma^2)^{-1}\|\mu_0\|_2^2 - \frac{1}{2}(\sigma_1^2+\sigma^2)^{-1}\|\mu_1\|_2^2 + \frac{1}{2}d\log\left(\frac{\sigma_0^2+\sigma^2}{\sigma_1^2+\sigma^2}\right) + \log\left(\frac{p}{1-p}\right).\end{aligned}$$

Since $\tilde{\Sigma}_s = (\sigma_s^2+\sigma^2)I$, $\tilde{\Sigma}_s^{1/2} = \sqrt{\sigma_s^2+\sigma^2}I$, and hence (28) simplifies to

$$M_0 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_1^2+\sigma^2}\right) + c + b^{\mathrm{T}}\mu_0 + \mu_0^{\mathrm{T}}A\mu_0 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_1^2+\sigma^2}\right) + c + \frac{2(\sigma_0^2+\sigma^2)\mu_1^{\mathrm{T}}\mu_0 - (\sigma_1^2+\sigma_0^2+2\sigma^2)\|\mu_0\|_2^2}{2(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}, \tag{155}$$

$$M_1 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2+\sigma^2}\right) + c + b^{\mathrm{T}}\mu_1 + \mu_1^{\mathrm{T}}A\mu_1 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2+\sigma^2}\right) + c + \frac{(\sigma_1^2+\sigma_0^2+2\sigma^2)\|\mu_1\|_2^2 - 2(\sigma_1^2+\sigma^2)\mu_1^{\mathrm{T}}\mu_0}{2(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}, \tag{156}$$

for $S=0$ and $S=1$, respectively. We also have that

$$\begin{aligned}u^{(0)} &= \tilde{\Sigma}_0^{1/2}(b + 2A\mu_0)\\ &= \frac{\sqrt{\sigma_0^2+\sigma^2}}{\sigma_1^2+\sigma^2}(\mu_1-\mu_0), &(157)\\ u^{(1)} &= \tilde{\Sigma}_1^{1/2}(b + 2A\mu_1)\\ &= \frac{\sqrt{\sigma_1^2+\sigma^2}}{\sigma_0^2+\sigma^2}(\mu_1-\mu_0). &(158)\end{aligned}$$

Therefore, (29) simplifies to

$$V_0 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_1^2+\sigma^2}\right)^2 + \frac{\sigma_0^2+\sigma^2}{(\sigma_1^2+\sigma^2)^2}\|\mu_1-\mu_0\|_2^2 = \frac{d(\sigma_1^2-\sigma_0^2)^2+2(\sigma_0^2+\sigma^2)\|\mu_1-\mu_0\|_2^2}{2(\sigma_1^2+\sigma^2)^2}, \tag{159}$$

$$V_1 = \frac{d}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2+\sigma^2}\right)^2 + \frac{\sigma_1^2+\sigma^2}{(\sigma_0^2+\sigma^2)^2}\|\mu_1-\mu_0\|_2^2 = \frac{d(\sigma_1^2-\sigma_0^2)^2+2(\sigma_1^2+\sigma^2)\|\mu_1-\mu_0\|_2^2}{2(\sigma_0^2+\sigma^2)^2}, \tag{160}$$

for $S=0$ and $S=1$, respectively. Substituting (155), (156), (159), and (160) into (31) yields

$$\mathrm{Var}(\theta^\sigma(X^\sigma)) = p\left(\frac{d(\sigma_1^2-\sigma_0^2)^2+2(\sigma_1^2+\sigma^2)\|\mu_1-\mu_0\|_2^2}{2(\sigma_0^2+\sigma^2)^2}\right) + (1-p)\left(\frac{d(\sigma_1^2-\sigma_0^2)^2+2(\sigma_0^2+\sigma^2)\|\mu_1-\mu_0\|_2^2}{2(\sigma_1^2+\sigma^2)^2}\right) \tag{161}$$

$$+ p(1-p)\left(\frac{d(\sigma_1^2-\sigma_0^2)^2+(\sigma_1^2+\sigma_0^2+2\sigma^2)\|\mu_1-\mu_0\|_2^2}{2(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}\right)^2. \tag{162}$$

Making appropriate substitutions, (32) simplifies to

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma),X^\sigma) &= a\left[2p(\sigma_1^2+\sigma^2)\mu_1^{\mathrm{T}} + 2(1-p)(\sigma_0^2+\sigma^2)\mu_0^{\mathrm{T}} + p(1-p)(d(\sigma_1^2-\sigma_0^2)+\|\mu_1\|_2^2-\|\mu_0\|_2^2)(\mu_1-\mu_0)^{\mathrm{T}}\right] \\ &\quad + b^{\mathrm{T}}\left[\alpha I + p(1-p)(\mu_1-\mu_0)(\mu_1-\mu_0)^{\mathrm{T}}\right] &(163)\\ &= \gamma_1\mu_1^{\mathrm{T}} + \gamma_0\mu_0^{\mathrm{T}}. &(164)\end{aligned}$$

where

$$\alpha \coloneqq p\sigma_1^2 + (1-p)\sigma_0^2 + \sigma^2 \tag{165}$$

and

$$\begin{aligned}\gamma_1 \coloneqq a\left[2p(\sigma_1^2+\sigma^2)+p(1-p)(d(\sigma_1^2-\sigma_0^2)+\|\mu_1\|_2^2-\|\mu_0\|_2^2)\right] + \frac{\alpha}{\sigma_1^2+\sigma^2} \\ + p(1-p)\frac{(\sigma_0^2+\sigma^2)\|\mu_1\|_2^2-(\sigma_1^2+\sigma_0^2+2\sigma^2)\mu_1^{\mathrm{T}}\mu_0+(\sigma_1^2+\sigma^2)\|\mu_0\|_2^2}{(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}, &(166)\\ \gamma_0 \coloneqq a\left[2(1-p)(\sigma_0^2+\sigma^2)-p(1-p)(d(\sigma_1^2-\sigma_0^2)+\|\mu_1\|_2^2-\|\mu_0\|_2^2)\right] - \frac{\alpha}{\sigma_0^2+\sigma^2} \\ - p(1-p)\frac{(\sigma_0^2+\sigma^2)\|\mu_1\|_2^2-(\sigma_1^2+\sigma_0^2+2\sigma^2)\mu_1^{\mathrm{T}}\mu_0+(\sigma_1^2+\sigma^2)\|\mu_0\|_2^2}{(\sigma_1^2+\sigma^2)(\sigma_0^2+\sigma^2)}. &(167)\end{aligned}$$

Again making the appropriate substitutions, (33) reduces to

$$\mathrm{Var}(X^\sigma) = \alpha I + p(1-p)(\mu_1-\mu_0)(\mu_1-\mu_0)^{\mathrm{T}}, \tag{168}$$

for α defined in (165). Using the Sherman-Morrison formula, we can compute $\mathrm{Var}^{-1}(X^\sigma)$ as follows:

$$\mathrm{Var}^{-1}(X^\sigma) = \alpha^{-1}I - \frac{\alpha^{-1}p(1-p)(\mu_1-\mu_0)(\mu_1-\mu_0)^{\mathrm{T}}}{\alpha+p(1-p)\|\mu_1-\mu_0\|_2^2}. \tag{169}$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(\theta^\sigma(X^\sigma),X^\sigma)\mathrm{Var}^{-1}(X^\sigma)\mathrm{Cov}(X^\sigma,\theta^\sigma(X^\sigma)) &= \alpha^{-1}(\gamma_1^2\|\mu_1\|_2^2+2\gamma_1\gamma_0\mu_1^{\mathrm{T}}\mu_0+\gamma_0^2\|\mu_0\|_2^2) \\ &\quad - \frac{\alpha^{-1}p(1-p)(\gamma_1\|\mu_1\|_2^2+(\gamma_0-\gamma_1)\mu_1^{\mathrm{T}}\mu_0-\gamma_0\|\mu_0\|_2^2)^2}{\alpha+p(1-p)\|\mu_1-\mu_0\|_2^2}, &(170)\end{aligned}$$

and hence

$$\mathrm{Var}(\theta^\sigma(X^\sigma)) - \mathrm{Cov}(\theta^\sigma(X^\sigma),X^\sigma)\mathrm{Var}^{-1}(X^\sigma)\mathrm{Cov}(X^\sigma,\theta^\sigma(X^\sigma)) = \frac{(\sigma_1^2-\sigma_0^2)^2(q_1+q_2\sigma^2+q_3\sigma^4+2d\sigma^6)}{4(r_1+r_2\sigma^2+r_3\sigma^4+r_4\sigma^6+r_5\sigma^8+\sigma^{10})}, \tag{171}$$

where

$$
\begin{aligned}
q_1 &= \|\mu_1-\mu_0\|_2^4\left(p^2(1-p)\sigma_1^2+p(1-p)^2\sigma_0^2\right)-d^2p^3(\sigma_1^2-\sigma_0^2)^3+2p(1-p)(2+d)\sigma_0^2\sigma_1^2\|\mu_1-\mu_0\|_2^2\\
&\quad+p^2\left(d(5d-2)\sigma_0^4\sigma_1^2-2d(2d+1)\sigma_0^2\sigma_1^4+d(2+d)\sigma_1^6-2d(d-1)\sigma_0^6\right)\\
&\quad+p\left(d(d-4)\sigma_0^6-2d(d-1)\sigma_0^4\sigma_1^2+d(2+d)\sigma_0^2\sigma_1^4\right)+2d\sigma_0^6,\\
q_2 &= p(1-p)\|\mu_1-\mu_0\|_2^2\left(\|\mu_1-\mu_0\|_2^2+2(2+d)(\sigma_0^2+\sigma_1^2)\right)-p^2d(d-4)(\sigma_1^2-\sigma_0^2)^2\\
&\quad+p\left(d(d-10)\sigma_0^4-2d(d-4)\sigma_0^2\sigma_1^2+d(2+d)\sigma_1^4\right)+6d\sigma_0^4,\\
q_3 &= 2p(1-p)(2+d)\|\mu_1-\mu_0\|_2^2+6d(p\sigma_1^2+(1-p)\sigma_0^2),\\
r_1 &= \sigma_0^4\sigma_1^4\left(p(1-p)\|\mu_1-\mu_0\|_2^2+p\sigma_1^2+(1-p)\sigma_0^2\right),\\
r_2 &= \sigma_0^2\sigma_1^2\left(2p(1-p)(\sigma_1^2+\sigma_0^2)\|\mu_1-\mu_0\|_2^2+2p\sigma_1^4+2(1-p)\sigma_0^4+3\sigma_0^2\sigma_1^2\right),\\
r_3 &= p(1-p)(\sigma_0^4+4\sigma_0^2\sigma_1^2+\sigma_1^4)\|\mu_1-\mu_0\|_2^2+p\sigma_1^6+3(1+p)\sigma_0^2\sigma_1^4+3(2-p)\sigma_0^4\sigma_1^2+(1-p)\sigma_0^6,\\
r_4 &= 2p(1-p)(\sigma_1^2+\sigma_0^2)\|\mu_1-\mu_0\|_2^2+(2p+1)\sigma_1^4+(3-2p)\sigma_0^4+6\sigma_0^2\sigma_1^2,\\
r_5 &= p(1-p)\|\mu_1-\mu_0\|_2^2+(3-p)\sigma_0^2+(2+p)\sigma_1^2.
\end{aligned}
$$