Appendices of "Intentional Control of Type I Error over Unconscious Data Distortion: a Neyman-Pearson Approach to Text Classification"

A. PROOFS

A.1 Proof of Theorem 1

Proof. Recall that the (classical) oracle classifier regarding the pre-distortion population is $h^*(x) = \mathbb{1}(\eta(x) > 1/2)$, where the regression function $\eta(x) = \mathbb{E}(Y|X = x)$ can be calculated as

$$\eta(x) = \frac{\pi_1 f_1(x)/f_0(x)}{\pi_1 f_1(x)/f_0(x) + \pi_0}\,.$$

Therefore, $h^*(x) = \mathbb{1}\left(\frac{f_1(x)}{f_0(x)} > \frac{\pi_0}{\pi_1}\right)$. When distortion with rates β_0 and β_1 is applied to class 0 and class 1 respectively, the class proportions become $\pi_0^{(\beta_0,\beta_1)}$ and $\pi_1^{(\beta_0,\beta_1)}$ which are defined as

$$\pi_0^{(\beta_0,\beta_1)} = \frac{(1-\beta_0)\pi_0}{(1-\beta_0)\pi_0 + (1-\beta_1)\pi_1}\,,$$
$$\pi_1^{(\beta_0,\beta_1)} = \frac{(1-\beta_1)\pi_1}{(1-\beta_0)\pi_0 + (1-\beta_1)\pi_1}\,,$$

while class conditional densities remain f_0 and f_1. Then, the oracle classifier regarding the post-distortion population is to replace π_0 and π_1 in h^* by $\pi_0^{(\beta_0,\beta_1)}$ and $\pi_1^{(\beta_0,\beta_1)}$ respectively:

$$h^*_{(\beta_0,\beta_1)}(x) = \mathbb{1}\left(\frac{f_1(x)}{f_0(x)} > \frac{\pi_0^{(\beta_0,\beta_1)}}{\pi_1^{(\beta_0,\beta_1)}}\right) = \mathbb{1}\left(\frac{f_1(x)}{f_0(x)} > \frac{1-\beta_0}{1-\beta_1}\cdot\frac{\pi_0}{\pi_1}\right)\,.$$

□

A.2 Proof of Theorem 2

Proof. The constrained optimization program (4) in the main text that defines ϕ^*_α does not involve the class priors $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$, so ϕ^*_α does not depend on π_0 or π_1. Now suppose distortion with rates β_0 and β_1 is imposed on class 0 and class 1 respectively, then the post-distortion population have class 0 proportion $[(1-\beta_0)\pi_0]/[(1-\beta_0)\pi_0 + (1-\beta_1)\pi_1]$ and class 1 proportion $[(1-\beta_1)\pi_1]/[(1-\beta_0)\pi_0 + (1-\beta_1)\pi_1]$, while keeping the distributions of $X|(Y = 0)$ and $X|(Y = 1)$ unchanged. Since distortion at rates β_0 and β_1 only changes class proportion, which NP oracle does not depend upon, the NP oracle is invariant under distortion. $\square$

B. COST-SENSITIVE (CS) LEARNING

An insight from studying the classical classification paradigm is that the relative size of classification errors comes largely from the relative weights placed on type I and type II errors in the objective function. So a natural candidate to adjust classification errors is to change the weights. This is the so-called cost-sensitive (CS) learning paradigm, in which users impose costs C_0 and C_1 to type I and type II errors, respectively. On the population level, instead of minimizing the overall classification error $R(\cdot)$, one minimizes the CS learning objective:

$$\min_h R^c(h) := C_0\pi_0 R_0(h) + C_1\pi_1 R_1(h)\,, \tag{A.1}$$

or the following variant of (A.1):

$$\min_h R^{\bar{c}}(h) := C_0 R_0(h) + C_1 R_1(h)\,. \tag{A.2}$$

Then, the CS oracle classifier h^{c*} under the cost-sensitive learning paradigm (A.1) can be calculated by

$$h^{c*}(x) = \mathbb{1}\left(\frac{f_1(x)}{f_0(x)} > \frac{C_0}{C_1}\cdot\frac{\pi_0}{\pi_1}\right)\,,$$

and the CS oracle $h^{\bar{c}*}$ under (A.2) can be calculated by

$$h^{\bar{c}*}(x) = \mathbb{1}\left(\frac{f_1(x)}{f_0(x)} > \frac{C_0}{C_1}\right)\,.$$

Similar to its counterpart in the classical paradigm, the post-distortion CS oracle classifier is different from the pre-distortion CS oracle, and the pre-distortion CS oracle cannot be recovered in view of an unknown distortion scheme. Lemma 1 follows from arguments similar to the proof of Theorem 1 in the main text.

Lemma 1. *Suppose that $X|(Y=0)$ and $X|(Y=1)$ have probability density functions f_0 and f_1, and that class priors are π_0 and π_1 respectively. Let β_0 and β_1 be the distortion rates of class 0 and class 1 respectively. Then, the oracle classifier under the cost-sensitive learning paradigm (A.1) regarding the post-distortion population is*

$$h^{c*}_{(\beta_0,\beta_1)}(x) = \mathbb{I}\left(\frac{f_1(x)}{f_0(x)} > \frac{1-\beta_0}{1-\beta_1}\cdot\frac{C_0}{C_1}\cdot\frac{\pi_0}{\pi_1}\right).$$

Similarly, the oracle classifier under the paradigm (A.2) regarding the post-distortion population is

$$h^{\bar{c}*}_{(\beta_0,\beta_1)}(x) = \mathbb{I}\left(\frac{f_1(x)}{f_0(x)} > \frac{1-\beta_0}{1-\beta_1}\cdot\frac{C_0}{C_1}\right).$$

Lemma 1 implies that even if we have the entire post-distortion population, we can only mimic $h^{c*}_{(\beta_0,\beta_1)}$ or $h^{\bar{c}*}_{(\beta_0,\beta_1)}$. However, unless β_0 and β_1 are known or estimable, there is no hope to mimic h^{c*} or $h^{\bar{c}*}$.

C. ORACLE CLASSIFIERS WHEN WE RELAX THE FIXED CLASS CONDITIONAL DENSITIES ASSUMPTION

Proposition 1. *Suppose that pre-distortion, $X|(Y=0)$ and $X|(Y=1)$ have probability density functions f_0 and f_1, and that class priors are $\pi_0 = \mathbb{P}(Y=0)$ and $\pi_1 = \mathbb{P}(Y=1)$. Let β_0 and β_1 be the distortion rates of class 0 and class 1 respectively. Further suppose that the post-distortion class conditional densities of features are f_0' and f_1'. Then, the classical oracle classifier regarding the pre-distortion population is*

$$h^*(x) = \mathbb{I}\left(\frac{f_1(x)}{f_0(x)} > \frac{\pi_0}{\pi_1}\right),$$

and that regarding the post-distortion population is

$$h^{*\prime}_{(\beta_0,\beta_1)}(x) = \mathbb{I}\left(\frac{f_1'(x)}{f_0'(x)} > \frac{1-\beta_0}{1-\beta_1}\cdot\frac{\pi_0}{\pi_1}\right).$$

The proof is omitted due to its similarity to that for Theorem 1 in the main text. Note that when $f_1'/f_0' = f_1/f_0$, that is when the ratio of class conditional densities of features is preserved under data distortion, the post-distortion classical oracle classifier $h^{*'}_{(\beta_0,\beta_1)}(x)$ reduces to $h^{*}_{(\beta_0,\beta_1)}(x)$ in Theorem 1, even if the class conditional densities themselves are changed. On the other hand, without assuming any relations between pre and post distortion feature distributions, f_1/f_0 cannot be recovered.

The invariance property (Theorem 2 in the main text) of Neyman-Pearson (NP) oracle classifiers no longer holds in general when the class conditional densities of features are different pre and post distortion. The next proposition illustrates sufficient and necessary conditions under which this invariance property does hold for a fixed α.

Proposition 2. *Denote pre-distortion distributions of $X|(Y=0)$ and $X|(Y=1)$ by f_0 and f_1 and those post-distortion by f_0' and f_1'. When $f_1'/f_0' = a\cdot(f_1/f_0)$ and*

$$a\cdot\min\{C\in\mathbb{R}:\mathbb{P}_{f_0}(f_1(X)/f_0(X)>C)\leq\alpha\} = \min\{C\in\mathbb{R}:\mathbb{P}_{f_0'}(f_1'(X)/f_0'(X)>C)\leq\alpha\}\,,$$

for some $a>0$, the NP oracle classifier ϕ^*_α defined in (4) in the main text is invariant under distortion at various rates β_0 (on class 0) and β_1 (on class 1), regardless of whether pre-distortion classes are balanced. Moreover, these conditions are also necessary for the invariance property.

Proof. From the NP Lemma, it is easy to see that the two conditions are sufficient for the invariance property of the NP oracles. For the necessary part, again by the NP lemma, the NP oracles pre and post distortion can be written respectively as

$$\phi^*_\alpha(x) = \mathbb{1}(f_1(x)/f_0(x)>C_\alpha),\ \text{and}\ \phi^{*'}_\alpha(x) = \mathbb{1}(f_1'(x)/f_0'(x)>C_\alpha')\,,$$

for some constants C_α and C_α' as determined in the NP Lemma. In other words,

$$C_\alpha = \min\{C\in\mathbb{R}:\mathbb{P}_{f_0}(f_1(X)/f_0(X)>C)\leq\alpha\}\,,$$

$$C_\alpha' = \min\{C\in\mathbb{R}:\mathbb{P}_{f_0'}(f_1'(X)/f_0'(X)>C)\leq\alpha\}\,.$$

Since C_α and C'_α are constants, to have $\phi^*_\alpha(x) = \phi^{*'}_\alpha(x)$, it is necessary to have $f'_1/f'_0 = a\cdot(f_1/f_0)$ for some positive constants a, and this further demands $C'_\alpha = a \cdot C_\alpha$. $\square$

Note that in general, the constant a in Proposition 2 depends on α. In the following, we demonstrate that within certain distribution classes, the more general condition in Proposition 2 falls back to the special case of unchanged class conditional feature distributions, while in others, there are $a \neq 1$ cases where class conditional feature distributions are different pre and post distortion.

Case I: Exponential Distribution Assume that $f_0(x) = \lambda_0 e^{-\lambda_0 x}$, $f_1(x) = \lambda_1 e^{-\lambda_1 x}$; $f'_0(x) = \lambda'_0 e^{-\lambda'_0 x}$, $f'_1(x) = \lambda'_1 e^{-\lambda'_1 x}$, where $x > 0$. For identifiability concern, let us assume $\lambda_0 < \lambda_1$, $\lambda'_0 < \lambda'_1$. Then,

$$\frac{f_1(x)}{f_0(x)} = \frac{\lambda_1}{\lambda_0} e^{-(\lambda_1 - \lambda_0)x}\,,$$

and

$$\frac{f'_1(x)}{f'_0(x)} = \frac{\lambda'_1}{\lambda'_0} e^{-(\lambda'_1 - \lambda'_0)x}\,.$$

When we demand

$$\frac{f'_1(x)}{f'_0(x)} = a \cdot \frac{f_1(x)}{f_0(x)} \qquad \forall x\,,$$

it follows that

$$\lambda_1 - \lambda_0 = \lambda'_1 - \lambda'_0\,, \tag{A.3}$$

and

$$\frac{\lambda'_1}{\lambda'_0} = a \cdot \frac{\lambda_1}{\lambda_0}\,. \tag{A.4}$$

Note that

$$
\begin{aligned}
P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C\right) &= P_{f_0}\left(\frac{\lambda_1}{\lambda_0}e^{-(\lambda_1-\lambda_0)X} > C\right) \\
&= P_{f_0}\left(e^{-(\lambda_1-\lambda_0)X} > \frac{\lambda_0}{\lambda_1}C\right) \\
&= P_{f_0}\left(X < -\frac{1}{\lambda_1-\lambda_0}\ln\left(\frac{\lambda_0}{\lambda_1}C\right)\right) \\
&= 1 - \exp\left\{-\lambda_0 \cdot \left[-\frac{1}{\lambda_1-\lambda_0}\ln\left(\frac{\lambda_0}{\lambda_1}C\right)\right]\right\} \\
&= 1 - \left(\frac{\lambda_0}{\lambda_1}C\right)^{\frac{\lambda_0}{\lambda_1-\lambda_0}} .
\end{aligned}
$$

To choose the minimum C such that $P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C\right) \leq \alpha$, we get

$$
C_\alpha = \frac{\lambda_1}{\lambda_0}(1-\alpha)^{\frac{\lambda_1-\lambda_0}{\lambda_0}} .
$$

Similarly,

$$
C'_\alpha = \frac{\lambda'_1}{\lambda'_0}(1-\alpha)^{\frac{\lambda'_1-\lambda'_0}{\lambda'_0}} .
$$

Then the condition $a \cdot C_\alpha = C'_\alpha$ implies that

$$
a \cdot \frac{\lambda_1}{\lambda_0}(1-\alpha)^{\frac{\lambda_1-\lambda_0}{\lambda_0}} = \frac{\lambda'_1}{\lambda'_0}(1-\alpha)^{\frac{\lambda'_1-\lambda'_0}{\lambda'_0}} . \tag{A.5}
$$

For any given $0 < \alpha < 1$, combining three equations (A.3), (A.4) and (A.5) implies that

$$
(1-\alpha)^{\frac{1}{\lambda_0}} = (1-\alpha)^{\frac{1}{\lambda'_0}} ,
$$

which implies that $\lambda_0 = \lambda'_0$. And then, $\lambda_1 = \lambda'_1$ and $a = 1$. Therefore, we have shown that when the class conditional feature distributions are restricted to the exponential distributions, the invariant property only occurs when $f_0 = f'_0$ and $f_1 = f'_1$.

Case II: Gaussian Distribution Assume that $f_0 : N(\mu_0, \sigma^2)$, $f_1 : N(\mu_1, \sigma^2)$, $f'_0 : N(\mu'_0, \sigma'^2)$, and $f_1' : N(\mu_1', \sigma'^2)$, where $\mu_0 < \mu_1$, $\mu_0' < \mu_1'$, and $\sigma \neq \sigma'$. Then,

$$\frac{f_1(x)}{f_0(x)} = \exp\left\{\frac{2(\mu_1 - \mu_0)x + \mu_0^2 - \mu_1^2}{2\sigma^2}\right\}$$

and

$$\frac{f_1'(x)}{f_0'(x)} = \exp\left\{\frac{2(\mu_1' - \mu_0')x + \mu_0'^2 - \mu_1'^2}{2\sigma'^2}\right\} .$$

To obtain

$$\frac{f_1'(x)}{f_0'(x)} = a \cdot \frac{f_1(x)}{f_0(x)} ,$$

the parameters $\mu_0, \mu_1, \sigma, \mu_0', \mu_1', \sigma', a$ must satisfy

$$\frac{2(\mu_1 - \mu_0)}{2\sigma^2} = \frac{2(\mu_1' - \mu_0')}{2\sigma'^2} , \tag{A.6}$$

and

$$a = \exp\left\{\frac{\mu_0'^2 - \mu_1'^2}{2\sigma'^2} - \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}\right\} .$$

Furthermore, denote by $\Phi(\cdot)$ the cumulative distribution function of standard normal distribution, $C_\alpha = \min_C \left\{C \in R : P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C\right) \leq \alpha\right\}$ and $C_\alpha' = \min_C \left\{C \in R : P_{f_0'}\left(\frac{f_1'(X)}{f_0'(X)} > C\right) \leq \alpha\right\}$.

$$\begin{aligned} P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C\right) &= P_{f_0}\left(\exp\left\{\frac{2(\mu_1 - \mu_0)X + \mu_0^2 - \mu_1^2}{2\sigma^2}\right\} > C\right) \\ &= P_{f_0}\left(2(\mu_1 - \mu_0)X + \mu_0^2 - \mu_1^2 > 2\sigma^2 \ln C\right) \\ &= P_{f_0}\left(X > \frac{2\sigma^2 \ln C + \mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)}\right) \\ &= P_{f_0}\left(\frac{X - \mu_0}{\sigma} > \frac{2\sigma^2 \ln C + (\mu_1 - \mu_0)^2}{2(\mu_1 - \mu_0)\sigma}\right) . \end{aligned}$$

Based on $P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C\right) \leq \alpha$, we get

$$\Phi^{-1}(1 - \alpha) \leq \frac{2\sigma^2 \ln C + (\mu_1 - \mu_0)^2}{2(\mu_1 - \mu_0)\sigma} ,$$

where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$, that is,

$$C \geq \exp\left\{\frac{2\sigma(\mu_1-\mu_0)\Phi^{-1}(1-\alpha)-(\mu_1-\mu_0)^2}{2\sigma^2}\right\}.$$

Therefore,

$$C_\alpha = \exp\left\{\frac{2\sigma(\mu_1-\mu_0)\Phi^{-1}(1-\alpha)-(\mu_1-\mu_0)^2}{2\sigma^2}\right\}.$$

Similarly,

$$C'_\alpha = \exp\left\{\frac{2\sigma'(\mu'_1-\mu'_0)\Phi^{-1}(1-\alpha)-(\mu'_1-\mu'_0)^2}{2\sigma'^2}\right\}.$$

From the relationship $a \cdot C_\alpha = C'_\alpha$, we can obtain

$$\begin{aligned} &\frac{\mu_0'^2-\mu_1'^2}{2\sigma'^2}-\frac{\mu_0^2-\mu_1^2}{2\sigma^2}+\frac{2\sigma(\mu_1-\mu_0)\Phi^{-1}(1-\alpha)-(\mu_1-\mu_0)^2}{2\sigma^2} \\ = &\frac{2\sigma'(\mu'_1-\mu'_0)\Phi^{-1}(1-\alpha)-(\mu'_1-\mu'_0)^2}{2\sigma'^2}, \end{aligned} \tag{A.7}$$

i.e.,

$$\begin{aligned} &\frac{\mu_0'^2-\mu_1'^2}{2\sigma'^2}+\frac{(\mu'_1-\mu'_0)^2}{2\sigma'^2}-\frac{\mu_0^2-\mu_1^2}{2\sigma^2}-\frac{(\mu_1-\mu_0)^2}{2\sigma^2} \\ = &\frac{(\mu'_1-\mu'_0)\Phi^{-1}(1-\alpha)}{\sigma'}-\frac{(\mu_1-\mu_0)\Phi^{-1}(1-\alpha)}{\sigma}, \end{aligned}$$

which is equivalent to,

$$\frac{\mu'_0(\mu'_0-\mu'_1)}{\sigma'^2}-\frac{\mu_0(\mu_0-\mu_1)}{\sigma^2}=\left[\frac{(\mu'_1-\mu'_0)}{\sigma'}-\frac{(\mu_1-\mu_0)}{\sigma}\right]\Phi^{-1}(1-\alpha). \tag{A.8}$$

From equation (A.6)

$$\frac{\mu'_1-\mu'_0}{\sigma'}=\frac{\sigma'}{\sigma^2}(\mu_1-\mu_0). \tag{A.9}$$

Putting (A.9) into (A.8),

$$\frac{(\mu_0-\mu_1)}{\sigma^2}(\mu'_0-\mu_0)=\left[\frac{\sigma'}{\sigma^2}(\mu_1-\mu_0)-\frac{(\mu_1-\mu_0)}{\sigma}\right]\Phi^{-1}(1-\alpha),$$

that is,

$$\Phi^{-1}(1-\alpha) = \frac{\mu_0 - \mu_0'}{\sigma' - \sigma}\,.$$

Putting the above arguments together, we have shown that under Gaussian distributions, for a given $\alpha \in (0,1)$, the invariance property is satisfied precisely when

$$\frac{(\mu_1 - \mu_0)}{\sigma^2} = \frac{(\mu_1' - \mu_0')}{\sigma'^2}\,,$$

$$\Phi^{-1}(1-\alpha) = \frac{\mu_0 - \mu_0'}{\sigma' - \sigma}\,,$$

and

$$a = \exp\left\{\frac{\mu_0'^2 - \mu_1'^2}{2\sigma'^2} - \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}\right\}\,.$$

Example of Case II: Let $f_0 : N(0,2^2)$, $f_1 : N(1,2^2)$ and $f_0' : N(-4,4^2)$, and $f_1' : N(0,4^2)$. We show that when $\alpha = 0.023$, the invariant property holds. First, it is easy to check that the above three equations hold with these density specifications and the choice of α. In the following, we provide an alternative direct proof.

Note that

$$\frac{f_1(x)}{f_0(x)} = \exp\left\{\frac{2x-1}{8}\right\}\,,$$

and

$$\frac{f_1'(x)}{f_0'(x)} = \exp\left\{\frac{8x+16}{32}\right\}\,,$$

Hence,

$$\frac{f_1'(x)}{f_0'(x)} = \exp\left\{\frac{5}{8}\right\}\cdot\frac{f_1(x)}{f_0(x)}\,.$$

We can take $a = \exp\left\{\frac{5}{8}\right\}$. Let $\alpha = 0.023$. Then $\Phi^{-1}(1-\alpha) = 2$. We solve for C_α and C_α' from

$$P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > C_\alpha\right) = \alpha \quad \text{and} \quad P_{f_0'}\left(\frac{f_1'(X)}{f_0'(X)} > C_\alpha'\right) = \alpha\,.$$

That is,

$$P_{f_0}\left(\exp\left\{\frac{2X-1}{8}\right\} > C_\alpha\right) = \alpha \quad \text{and} \quad P_{f_0'}\left(\exp\left\{\frac{8X+16}{32}\right\} > C_\alpha'\right),$$

Or equivalently,

$$P_{f_0}\left(X > \frac{8\ln C_\alpha + 1}{2}\right) = \alpha \quad \text{and} \quad P_{f_0'}\left(X > 4\ln C_\alpha' - 2\right) = \alpha\,.$$

That is,

$$\frac{(8\ln C_\alpha + 1)/2}{2} = 2 \quad \text{and} \quad \frac{4\ln C_\alpha' - 2 - (-4)}{4} = 2,$$

which implies that

$$C_\alpha = \exp\left\{\frac{7}{8}\right\} \quad \text{and} \quad C_\alpha' = \exp\left\{\frac{3}{2}\right\}.$$

Obviously,

$$a \cdot C_\alpha = C_\alpha',$$

i.e.,

$$a \cdot \min_C \left\{C \in R : P_{f_0}\left(\frac{f_1}{f_0} > C\right) \le \alpha\right\} = \min_C \left\{C \in R : P_{f_0'}\left(\frac{f_1'}{f_0'} > C\right) \le \alpha\right\}.$$

Therefore, we have constructed a concrete NP oracle invariant example in which $f_0 \neq f_0'$ and $f_1 \neq f_1'$.

D. SPARSITY-INDUCING METHODS IN SELECTING MEANINGFUL TOPICS

Among the implemented methods, NP-sLDA performs the best in terms of power and it is a penalized sparsity-inducing method, which means it eliminates certain unimportant features as part of the classifier training process. In this section, we elaborate that such methods are effective in terms of selecting meaningful topics. In particular, we look at results from the first two random repetitions under Setting 1 in Section 4.5.2 (random seed being set and results are readily available online) with $K = 10$. In the first repetition, Table A1 displays the selected ten topics and it's obvious that only topics 4 and 10 are the strike-related topics. Following the common practice of NP umbrella algorithms, we randomly split the training data M times for training the scoring function and thresholds. Here we use $M = 7$, and the final classifier is a majority vote. Figure A1

topic 1	罢工	终于	学校	时间	一下	事件	彻底	哼哼	开	对
	strike	finally	school	time	a bit	event	complete	humph	open	right
	电话	集体	分钟	失望	胃	为了	好多	疑问	多少	忙
	phone	collective	minute	disappoint	stomach	for	many	question	how many	busy
topic 2	罢工	今天	上班	发生	年	上	发现	问题	种	衰
	strike	today	work	happen	year	go	discover	problem	type	decline
	回来	太阳	话	公交车	宿舍	冷	块	过节	东西	思考
	come back	Sun	words	bus	dorm	cold	block	festival	things	think
topic 3	人	让	说	吃	时候	事	罢课	过	哈哈	小
	people	let	speak	eat	time	thing	student strike	pass	haha	small
	里	妈妈	老师	今晚	去	很多	找	出门	最近	班
	inside	mom	teacher	tonight	go	many	find	go out	recent	class
topic 4	年	公司	员工	中	工人	工作	工资	最后	月	鄙视
	year	company	employee	within	worker	work	salary	finally	month	despise
	小时	后	广州	抗议	今日	知道	请	月日	要求	中国
	hour	after	Guangzhou	protest	today	know	please	month-date	request	China
topic 5	抓	狂	罢工	电脑	泪	早上	现在	抓狂	天气	回家
	clutch	crazy	strike	computer	tear	morning	now	go crazy	weather	go home
	想	潮湿	下午	结果	集	继续	部	修	人	委屈
	think	moist	afternoon	result	gather	continue	department	fix	people	be wronged
topic 6	罢工	去	做	次	能	生病	地	系	偷笑	没有
	strike	go	do	times	can	sick	ground	systems	smirk	without
	睡觉	鼻屎	挖	第一	今晚	怒	回	真的	叫	汗
	sleep	mucus	pick	first	tonight	angry	back	real	shout	sweat
topic 7	天	手机	还是	知道	能	竟然	突然	说	玩	这个
	day	cellphone	still	know	can	unexpectedly	suddenly	say	play	this
	出来	换	已经	点	郁闷	鼓掌	听	一下	真是	好不容易
	out	exchange	already	bit	depressed	applaud	listen	a bit	really	hard
topic 8	罢工	想	可怜	居然	买	发	明天	累	点	但是
	strike	think	pity	unexpectedly	buy	give	tomorrow	tired	bit	but
	星期	然后	休息	家里	半	悲伤	一直	本来	听说	心情
	week	therefore	rest	home	half	sad	always	originally	heard	mood
topic 9	罢工	草草	明天	可以	开始	好好	真的	新闻	爱	开
	strike	hastily	tomorrow	can	start	nicely	really	news	love	open
	心情	点	还有	刚刚	这个	之后	一定	为什么	晚	上午
	mood	a bit	also	just	this	after	must	why	evening	morning
topic 10	的士	汕头	出租车	司机	现在	车	罢工	打	下	辆
	taxi	Shantou	taxi	driver	now	car	strike	call	get off	vehicle
	营运	三	原因	政府	集体	今日	希望	四	市民	月日
	operate	three	reason	government	collective	today	hope	four	citizen	month-date

Table A1: top 20 keywords for the ten topics selected from repetition 1.

shows that, over the seven splits, NP-sLDA consistently selects only topics 4 and 10, and all the rest of the topics have corresponding coefficient 0. Similarly, in repetition 2, Table A2 shows that only topics 5 and 6 are the strike-related topics, and Figure A2 shows that NP-sLDA consistently selects topics 5 and 6 over the 7 splits. In summary, these sparsity-inducing methods, such as NP-sLDA, help select meaningful topics.

E. PROOF AND GENERALIZATION OF PROPOSITION 1 IN MAIN TEXT

Proposition 1 in the main text follows as a special case of the next Proposition. Proposition 3 below explores the relationship between type I error $R_0(\cdot)$, the distortion rate β_0 of class 0 and the class size ratio π_0/π_1 for the classical post-distortion oracle classifier $h^*_{\beta_0,\pi_0}$.

(Intercept)	2.526526	(Intercept)	2.592229	(Intercept)	2.811912	(Intercept)	2.067274
x1	.	x1	.	x1	.	x1	.
x2	.	x2	.	x2	.	x2	.
x3	.	x3	.	x3	.	x3	.
x4	-5.914467	x4	-7.982268	x4	-7.691559	x4	-4.295535
x5	.	x5	.	x5	.	x5	.
x6	.	x6	.	x6	.	x6	.
x7	.	x7	.	x7	.	x7	.
x8	.	x8	.	x8	.	x8	.
x9	.	x9	.	x9	.	x9	.
x10	-19.521468	x10	-17.877815	x10	-20.608779	x10	-16.470045

(Intercept)	2.286663	(Intercept)	3.21628	(Intercept)	2.438294
x1	.	x1	.	x1	.
x2	.	x2	.	x2	.
x3	.	x3	.	x3	.
x4	-5.279157	x4	-11.65932	x4	-7.726201
x5	.	x5	.	x5	.
x6	.	x6	.	x6	.
x7	.	x7	.	x7	.
x8	.	x8	.	x8	.
x9	.	x9	.	x9	.
x10	-17.663786	x10	-20.64825	x10	-16.653843

Figure A1: regression coefficients for the 7 splits in NP-sLDA, repetition 1.

topic 1	今天	天	可以	没有	开始	点	真的	日子	明天	能
	today	day	can	without	start	bit	really	day	tomorrow	can
	前	回家	还有	吃饭	吃	号	那些	地铁	哈哈	玩
	forward	go home	also	eat	eat	day	those	subway	haha	play
topic 2	罢工	上	上班	能	终于	抓狂	拿	小时	里	东西
	strike	go to	work	can	finally	go crazy	get	hour	inside	thing
	真是	三	为了	生活	之后	超级	只是	开心	觉得	对
	really	three	for	life	after	super	just	happy	feel	right
topic 3	罢工	去	系	做	今日	地	睡觉	后	起来	听
	strike	go	be	do	today	ground	sleep	after	get up	listen
	搞	过	怒	公交	求	人	甘	吃	街	说
	do	over	angry	public transportation	beg	people	willing	eat	street	speak
topic 4	想	人	让	说	罢课	累	发	衰	生病	找
	think	people	let	speak	student strike	tired	happen	decline	sick	find
	次	鄙视	很多	新	但是	哦	感冒	虽然	委屈	竟然
	time	despise	many	new	but	oh	a cold	although	be wronged	unexpectedly
topic 5	年	公司	月	员工	工人	工资	最后	月日	工作	还是
	year	company	month	employee	worker	salary	finally	month-date	work	still
	集体	第一	国际	次	要求	劳动	无法	机场	买	法国
	collective	first	international	time	demand	labor	unable	airport	buy	France
topic 6	罢工	的士	汕头	现在	出租车	司机	车	打	集体	辆
	strike	taxi	Shantou	now	taxi	driver	car	call	collective	vehicle
	下	广州	出门	已经	政府	事件	出	路	钱	问题
	get off	Guangzhou	go out	already	government	event	out	street	money	problem
topic 7	可怜	小	结果	偷笑	发现	昨天	今晚	早上	能	一直
	pity	small	result	smirk	find	yesterday	tonight	morning	can	always
	生病	竟然	郁闷	开	星期	罢工	三	今天	出来	结局
	sick	unexpectedly	depressed	open	week	strike	three	today	go out	end
topic 8	罢工	天	知道	天气	时候	挖	鼻屎	太阳	种	周
	strike	day	know	weather	time	pick	mucus	Sun	type	week
	今天	时间	电视	突然	奥特曼	好像	应该	全部	水	点
	today	time	TV	sudden	Ultraman	maybe	should	whole	water	bit
topic 9	罢工	抓	狂	泪	电脑	居然	今晚	鼓掌	泪泪	学校
	strike	clutch	crazy	tear	computer	unexpectedly	tonight	applaud	tear	school
	事	搞到	部	学生	结	啊啊	手机	明天	闹	闹钟
	thing	get	department	student	form	ah	cellphone	tomorrow	alarm	alarm clock
topic 10	罢工	手机	中	最近	哼哼	一下	哈哈	电梯	停播	玩
	strike	cellphone	within	recent	humph	a bit	haha	elevator	stop playing	play
	女	怒骂	分钟	时候	过	晚	深圳	第一	迟到	下班
	female	curse	minute	time	over	late	Shenzhen	first	late	off work

Table A2: top 20 keywords for the ten topics selected from repetition 2.

(Intercept)	2.890680	(Intercept)	2.541224	(Intercept)	2.588368	(Intercept)	2.523783
x1	.	x1	.	x1	.	x1	.
x2	.	x2	.	x2	.	x2	.
x3	.	x3	.	x3	.	x3	.
x4	.	x4	.	x4	.	x4	.
x5	-7.087097	x5	-7.949472	x5	-6.654200	x5	-6.538955
x6	-21.213795	x6	-17.069710	x6	-18.605443	x6	-18.279681
x7	.	x7	.	x7	.	x7	.
x8	.	x8	.	x8	.	x8	.
x9	.	x9	.	x9	.	x9	.
x10	.	x10	.	x10	.	x10	.

(Intercept)	2.682391	(Intercept)	2.450597	(Intercept)	2.917165
x1	.	x1	.	x1	.
x2	.	x2	.	x2	.
x3	.	x3	.	x3	.
x4	.	x4	.	x4	.
x5	-4.826610	x5	-5.872407	x5	-9.101548
x6	-21.347954	x6	-18.176870	x6	-19.616616
x7	.	x7	.	x7	.
x8	.	x8	.	x8	.
x9	.	x9	.	x9	.
x10	.	x10	.	x10	.

Figure A2: regression coefficients for the 7 splits in NP-sLDA, repetition 2.

Proposition 3. *Suppose probability densities of class* 0 $(X|Y=0)$ *and class* 1 $(X|Y=1)$ *follow distributions* $\mathcal{N}(\mu_0,\Sigma)$ *and* $\mathcal{N}(\mu_1,\Sigma)$ *respectively; class* 0 *composes* $\pi_0 \in (0,1)$ *proportion of the population and* $\beta_0 \in (0,1)$ *is the censorship rate of class* 0 *(i.e., the proportion of class* 0 *posts that were removed from some government censorship scheme). Suppose class 1 is not distorted (i.e.,* $\beta_1 = 0$*). Let* $h^*_{\beta_0,\pi_0}$ *be the classical oracle classifier in the post-distortion population. Then the type I error of* $h^*_{\beta_0,\pi_0}$ *(regarding either the pre-distortion or the post-distortion population) is calculated as :*

$$R_0(h^*_{\beta_0,\pi_0}) = \Phi\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right), \tag{A.10}$$

where $C = (\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0-\mu_1)$ *and* $p = \pi_0/(1-\pi_0)$*. Equation (A.10) implies that*

1. *Keeping* π_0 *fixed (hence* p *is fixed),* $R_0(h^*_{\beta_0,\pi_0})$ *is a monotone increasing function of the class* 0 *censorship rate* $\beta_0 \in (0,1)$*. Moreover, we have i). if* $pe^{3C/2} \leq 1$*,* $R_0(h^*_{\beta_0,\pi_0})$ *is a concave function of* $\beta_0 \in (0,1)$*; and ii). if* $pe^{3C/2} > 1$*,* $R_0(h^*_{\beta_0,\pi_0})$ *is a convex function of* β_0 *for* $\beta_0 \in \left(0, 1-\frac{1}{pe^{3C/2}}\right)$*, and a concave function for* $\beta_0 \in \left(1-\frac{1}{pe^{3C/2}}, 1\right)$*.*
2. *Keeping* β_0 *fixed,* $R_0(h^*_{\beta_0,\pi_0})$ *is a monotone decreasing function of the class ratio* $p = \pi_0/(1-\pi_0)$*. In other words, the larger the proportion of class* 0 *in the uncensored population, the smaller the type I error of* $h^*_{\beta_0,\pi_0}$*. Moreover,* $R_0(h^*_{\beta_0,\pi_0})$ *is a convex function of* p *for* $p > \frac{1}{(1-\beta_0)e^{3C/2}}$*, and it is a concave function of* p *for* $p \leq \frac{1}{(1-\beta_0)e^{3C/2}}$*.*

Proof. Since equation (2) in the main text is the decision boundary of $h^*_{\beta_0,\pi_0}$, we have

$$R_0(h^*_{\beta_0,\pi_0}) = P_{X\sim\mathcal{N}(\mu_0,\Sigma)}\left\{X^\top\Sigma^{-1}(\mu_0-\mu_1) - \frac{1}{2}(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0+\mu_1) + \log\left(\frac{(1-\beta_0)\pi_0}{\pi_1}\right) \leq 0\right\}.$$

For X in class 0, $X^\top\Sigma^{-1}(\mu_0-\mu_1) =: Z' \sim \mathcal{N}(\mu_0^\top\Sigma^{-1}(\mu_0-\mu_1), (\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0-\mu_1))$. Therefore,

$$\begin{aligned} R_0(h^*_{\beta_0,\pi_0}) &= P_{Z'\sim\mathcal{N}(\mu_0^\top\Sigma^{-1}(\mu_0-\mu_1),(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0-\mu_1))}\left\{Z' \leq \frac{1}{2}(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0+\mu_1) - \log\left(\frac{(1-\beta_0)\pi_0}{\pi_1}\right)\right\} \\ &= \Phi\left(\frac{-\frac{1}{2}(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0-\mu_1) - \log\left(\frac{(1-\beta_0)\pi_0}{\pi_1}\right)}{\sqrt{(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0-\mu_1)}}\right). \end{aligned}$$

Regarding part 1, for fixed π_0, let $f(\beta_0) = R_0(h^*_{\beta_0,\pi_0})$.

$$f'(\beta_0) = \phi\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{1}{\sqrt{C}(1-\beta_0)},$$

where $\phi(\cdot)$ is the probability density function of the standard normal random variable. This implies that for $\beta_0 \in (0,1)$, $f'(\cdot)$ is positive, so $R_0(h^*_{\beta_0,\pi_0})$ is a monotone increasing function of β_0 for fixed π_0. Taking the second derivative of f, we have

$$f''(\beta_0) = \phi'\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{1}{C(1-\beta_0)^2} + \phi\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{1}{\sqrt{C}(1-\beta_0)^2}.$$

Let $g(w) = \phi'(w) + \sqrt{C}\phi(w)$. Then

$$g(w) = \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}} \cdot (-w) + \frac{\sqrt{C}}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}.$$

Note that $g(w) > 0$ iff $w < \sqrt{C}$.

Therefore, $f''(\beta_0) > 0$ iff $g(\frac{-\frac{1}{2}C - \log((1-\beta_0)p)}{\sqrt{C}}) > 0$ iff $\frac{-\frac{1}{2}C - \log((1-\beta_0)p)}{\sqrt{C}} < \sqrt{C}$ iff $\beta_0 < 1 - \frac{1}{pe^{3C/2}}$. Similarly $f''(\beta_0) < 0$ iff $\beta_0 > 1 - \frac{1}{pe^{3C/2}}$.

Regarding part 2, for fixed β_0, let $k(p) = R_0(h^*_{\beta_0,\pi_0})$, then

$$k'(p) = \phi\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{-1}{\sqrt{C}p}.$$

Clearly, $k'(p) < 0$ for all $p > 0$.

$$k''(p) = \phi'\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{1}{Cp^2} + \phi\left(\frac{-\frac{1}{2}C - \log\left((1-\beta_0)p\right)}{\sqrt{C}}\right) \cdot \frac{1}{\sqrt{C}p^2}.$$

Note that $k''(p) > 0$ iff $\frac{-\frac{1}{2}C - \log((1-\beta_0)p)}{\sqrt{C}} < \sqrt{C}$ iff $p > \frac{1}{(1-\beta_0)e^{3C/2}}$. $\square$

The constant C can be considered as a measure of separability of the two classes. Note that when $p = 1$, that is when $\pi_0 = 1 - \pi_0 = 1/2$, if C is large (i.e., it is easy to separate the two classes), $1/(pe^{3C/2}) \approx 0$, then $R_0(h^*_{\beta_0,\pi_0})$ is a convex function of $\beta_0 \in (0,1)$. On the other hand, when C is so small (i.e., two classes are hard to separate) that $pe^{3C/2} \leq 1$, $R_0(h^*_{\beta_0,\pi_0})$ is a concave function of $\beta_0 \in (0,1)$.

F. NEYMAN-PEARSON LEMMA

The oracle classifier under the NP paradigm (NP oracle) arises from its close connection to the Neyman-Pearson Lemma in statistical hypothesis testing. Hypothesis testing bears strong resemblance to binary classification if we assume the following model. Let P_1 and P_0 be two *known* probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Assume that $Y \sim \text{Bern}(\zeta)$ for some $\zeta \in (0,1)$, and the conditional distribution of X given Y is P_Y. Given such a model, the goal of statistical hypothesis testing is to determine if we should reject the null hypothesis that X was generated from P_0. To this end, we construct a randomized test $\phi : \mathcal{X} \to [0,1]$ that rejects the null with probability $\phi(X)$. Two types of errors arise: type I error occurs when P_0 is rejected yet $X \sim P_0$, and type II error occurs when P_0 is not rejected yet $X \sim P_1$. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing ϕ that solves the following constrained optimization problem

$$\text{maximize } \mathbb{E}[\phi(X)|Y=1]\,, \text{ subject to } \mathbb{E}[\phi(X)|Y=0] \le \alpha\,,$$

where $\alpha \in (0,1)$ is the significance level of the test. A solution to this constrained optimization problem is called *a most powerful test* of level α. The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

Lemma 2 (Neyman-Pearson Lemma). *Let P_1 and P_0 be two probability measures with densities f_1 and f_0 respectively, and denote the density ratio as $r(x) = f_1(x)/f_0(x)$. For a given significance level α, let C_α be such that $P_0\{r(X) > C_\alpha\} \le \alpha$ and $P_0\{r(X) \ge C_\alpha\} \ge \alpha$. Then, the most powerful test of level α is*

$$\phi^*_\alpha(X) = \begin{cases} 1 & \text{if } r(X) > C_\alpha\,, \\ 0 & \text{if } r(X) < C_\alpha\,, \\ \frac{\alpha - P_0\{r(X)>C_\alpha\}}{P_0\{r(X)=C_\alpha\}} & \text{if } r(X) = C_\alpha\,. \end{cases}$$

Under mild continuity assumption, we take the *NP oracle classifier*

$$\phi^*_\alpha(x) \;=\; \mathbb{1}\{f_1(x)/f_0(x) > C_\alpha\} \;=\; \mathbb{1}\{r(x) > C_\alpha\}\,, \tag{A.11}$$

as our plug-in target for NP classification.