Locally Private Hypothesis Testing — Supplementary Material

A. Additional Claims

Proposition 14. *For any $a, b > 0$ we have $(a+b)^{2/3} \leq a^{2/3} + b^{2/3}$.*

Proof. Let $f(t) \stackrel{\text{def}}{=} (t+1)^{2/3} - t^{2/3} - 1$. Clearly, $f(0) = 0$. Moreover, $f'(t) = \frac{2}{3}\left((t+1)^{-1/3} - t^{1/3}\right)$. Since $t+1 > t > 0$ it follows that $(t+1)^{-1/3} < t^{-1/3}$ and so $f'(t) < 0$ for any $t \in (0, \infty)$. Therefore, for any $t > 0$ we have $f(t) < f(0) = 0$. Fix $a, b > 0$ and now we have:

$$0 > \left(\tfrac{a}{b} + 1\right)^{2/3} - \left(\tfrac{a}{b}\right)^{2/3} - 1 = \left(\tfrac{a+b}{b}\right)^{2/3} - \left(\tfrac{a}{b}\right)^{2/3} - 1$$

hence $a^{2/3} + b^{2/3} > (a+b)^{2/3}$. □

Proposition 15. *For any $a, b > 0$ we have $(a+b)^{3/2} = \Theta\left(a^{3/2} + b^{3/2}\right)$.*

Proof. Clearly, due to the non-negativity of a and b we have $(a+b)^{3/2} \leq (2\max\{a,b\})^{3/2} \leq \sqrt{8}(a^{3/2} + b^{3/2})$. Similarly, $a^{3/2} + b^{3/2} \leq 2(a+b)^{3/2}$. □

Claim 16. *Fix two constants $0 < \eta < \mu < 1$. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ be a collection of n vectors in $\mathbb{R}^d$ whose entries are generated iid and uniformly among $\{\mu - \eta, \mu + \eta\}$. If $n = \Omega(\frac{d^2 \log^2(d/\delta\eta)}{\eta^2})$ then for any unit-length vector $\boldsymbol{u} \in \mathbb{R}^d$ we have $\frac{1}{n}\sum_i (\boldsymbol{x}_i^\mathsf{T}\boldsymbol{u})^2 > \eta^2/3$.*

Proof. Denote $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ the collection of n vectors such that $\boldsymbol{y}_i = \frac{1}{\eta}(\boldsymbol{x}_i - \mu\mathbf{1})$. Therefore, for each $\boldsymbol{y}_i$, its coordinates are chosed iid an uniformly among $\{-1, 1\}$. Therefore $\mathrm{E}[\boldsymbol{y}_i] = \mathbf{0}$, and $\mathrm{E}[\boldsymbol{y}_i \boldsymbol{y}_i^\mathsf{T}] = I_{d\times d}$. As a $\frac{\eta}{12\sqrt{d}}$-cover of the unit-sphere in $\mathbb{R}^d$ contain $O(\eta^{-d\log(d)})$ points (see Vershynin (2010) for proof), standard Hoeffding-and-union bound yield that $\Pr\left[\exists \boldsymbol{u} \text{ in the cover s.t. } \left|\left(\frac{1}{n}\sum_i \boldsymbol{y}_i\right)^\mathsf{T}\boldsymbol{u} - 0\right| > \frac{\eta}{12\sqrt{d}}\right] \leq (\frac{1}{\eta})^{100d\log(d)} \cdot 2\exp(-2n\frac{\eta^2}{144d}) < \frac{\delta}{2}$. The triangle inequality thus assures us that for any unit-length vector in $\mathbb{R}^d$ we have $\left|\left(\frac{1}{n}\sum_i \boldsymbol{y}_i\right)^\mathsf{T}\boldsymbol{u}\right| \leq \frac{\eta}{6\sqrt{d}}$. Moreover, standard matrix-concentration results (Vershynin, 2010) on the spectrum of the matrix $\frac{1}{n}\sum_i \boldsymbol{y}_i\boldsymbol{y}_i^\mathsf{T}$ give that w.p. $1 - \frac{\delta}{2}$ we have that for any unit-length vector $\boldsymbol{u}$ it holds that

$$\boldsymbol{u}^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{y}_i\boldsymbol{y}_i^\mathsf{T} - I\right)\boldsymbol{u} = O\left(\frac{\sqrt{d} + \log(1/\delta)}{\sqrt{n}}\right) \leq \frac{1}{3}$$

by our choice of n.

Assume both events hold. As for each i we have $\boldsymbol{x}_i = \mu\mathbf{1} + \eta\boldsymbol{y}_i$ then it holds that $\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} = \mu^2 1_{d\times d} + \frac{\mu\eta}{n}(\sum_i \mathbf{1}\boldsymbol{y}_i^\mathsf{T} + \boldsymbol{y}_i\mathbf{1}^\mathsf{T}) + \frac{\eta^2}{n}\sum_i \boldsymbol{y}_i\boldsymbol{y}_i^\mathsf{T}$, thus for each unit-length $\boldsymbol{u}$ we have

$$\begin{aligned}\frac{1}{n}\sum_i (\boldsymbol{x}_i^\mathsf{T}\boldsymbol{u})^2 &= \mu^2(\mathbf{1}^\mathsf{T}\boldsymbol{u})^2 + 2\mu\eta(\mathbf{1}^\mathsf{T}\boldsymbol{u}) \cdot \left(\tfrac{1}{n}\sum_i \boldsymbol{y}_i\right)^\mathsf{T}\boldsymbol{u} \\ &\quad + \eta^2\boldsymbol{u}^\mathsf{T}\left(\tfrac{1}{n}\sum_i \boldsymbol{y}_i\boldsymbol{y}_i^\mathsf{T}\right)\boldsymbol{u} \\ &\geq 0 - 2\cdot 1\cdot\eta\sqrt{d}\cdot\tfrac{\eta}{6\sqrt{d}} + \eta^2\cdot\tfrac{2}{3} = \eta^2/3\end{aligned}$$

□

Proposition 17. *Let $\|\cdot\|$ be any norm satisfying $\|\boldsymbol{u} \otimes \boldsymbol{v}\| = \|\boldsymbol{u}\|\|\boldsymbol{v}\|$ (such as the L_p-norm for any $p \geq 1$). Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2$ be vectors whose norms are all bounded by some c. Then $\|\boldsymbol{x}_1 \otimes \boldsymbol{y}_1 - \boldsymbol{x}_2 \otimes \boldsymbol{y}_2\| \leq c\left(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|\right)$.*

Proof.

$$\begin{aligned}&\|\boldsymbol{x}_1 \otimes \boldsymbol{y}_1 - \boldsymbol{x}_2 \otimes \boldsymbol{y}_2\| \\ &= \|\boldsymbol{x}_1 \otimes \boldsymbol{y}_1 - \boldsymbol{x}_1 \otimes \boldsymbol{y}_2 + \boldsymbol{x}_1 \otimes \boldsymbol{y}_2 - \boldsymbol{x}_2 \otimes \boldsymbol{y}_2\| \\ &\leq \|\boldsymbol{x}_1 \otimes \boldsymbol{y}_1 - \boldsymbol{x}_1 \otimes \boldsymbol{y}_2\| + \|\boldsymbol{x}_1 \otimes \boldsymbol{y}_2 - \boldsymbol{x}_2 \otimes \boldsymbol{y}_2\| \\ &= \|\boldsymbol{x}_1\| \cdot \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| + \|\boldsymbol{y}_2\| \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \\ &\leq c\left(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|\right)\end{aligned}$$

□

B. Missing Proofs: Symmetric Scheme

Corollary 18 (Corollary 2 restated.)**.** *Let $\boldsymbol{q}^*$ be the $|\mathcal{S}|$-dimensional vector given by $\langle \frac{n_s}{n} \rangle$. Given that $|\mathcal{S}| \leq |\mathcal{X}|$, that G is a full-rank matrix satisfying $\|G\|_1 = 1$ and assuming that $\left(G^\dagger\boldsymbol{q}^* + \ker(G)\right) \cap H \neq \emptyset$, then any vector in H of the form $\boldsymbol{p}^* + \boldsymbol{u}$ where $\boldsymbol{p}^* = G^\dagger\boldsymbol{q}^*$ and $\boldsymbol{u} \in \ker(G)$ is an hypothesis that maximizes the likelihood of the given signals $(y_1, ..., y_n)$.*

Proof. Our goal is to find some $\boldsymbol{p} \in H$ which minimizes $f(\boldsymbol{p})$. Denoting $\boldsymbol{q}$ as the $|\mathcal{S}|$-dimensional vector such that $q(s) = \boldsymbol{g}_s^\mathsf{T}\boldsymbol{p}$, we note that G isn't just any linear transformation, but rather one that induces probability over the signals, and so $\boldsymbol{q}$ is a non-negative vector that sums to 1. We therefore convert the problem of minimizing our loss func-

tion into the following optimization problem

$$\begin{aligned}\min \phi(\boldsymbol{p},\boldsymbol{q}) &= -\sum_{s\in S} n_s \log(q(s)) \\ \text{subject to} \sum_s q(s) &= 1 \\ \forall s, q(s) &\geq 0 \\ \boldsymbol{q} &= G\boldsymbol{p} \\ \boldsymbol{p} &\in H\end{aligned}$$

Using Lagrange multipliers, it is easy to see that $\frac{\partial\phi}{\partial \boldsymbol{q}} = \langle \frac{-n_s}{q(s)} \rangle_{s\in\mathcal{S}}$ and that $\frac{\partial}{\partial \boldsymbol{q}}\left(\sum_{s\in\mathcal{S}} q(s) - 1\right) = \mathbf{1} = \frac{\partial}{\partial \boldsymbol{q}}\left(\boldsymbol{q} - G\boldsymbol{p} = 0\right)$ and so the minimizer is obtained when $\boldsymbol{q}$ equates all ratios $\frac{n_s}{q(s)} = \frac{n_{s'}}{q(s')}$ for all s, s', namely when $\boldsymbol{q} = \boldsymbol{q}^*$. Since we assume $G^\dagger \boldsymbol{q}^* + \ker(G)$ has a non-empty intersection with H, then let $\boldsymbol{p}$ be any hypothesis in H of the form $\boldsymbol{p}^* + \boldsymbol{u}$ where $\boldsymbol{u} \in \ker(G)$. We get that $(\boldsymbol{p}, \boldsymbol{q})$ is the minimizer of ϕ satisfying all constraints. By assumption, $\boldsymbol{p} \in H$. Due to the fact that G is full-rank and that $|\mathcal{S}| \leq |\mathcal{X}|$ we have that $G(\boldsymbol{p}^* + u) = G \cdot G^\dagger \boldsymbol{q}^* + \mathbf{0} = I \cdot \boldsymbol{q}^* = \boldsymbol{q}^*$, and by definition, $\boldsymbol{q}^*$ is a valid distribution vector (non-negative that sums to 1). □

Claim 19 (Claim 4)**.** *Given signals $y_1, ..., y_n$ generated using standard randomized response with parameter $\epsilon < 1$, we have that our log-loss function is $\Theta(\epsilon^2 \cdot \frac{\min_x\{n_x\}}{n})$-strongly convex.*

Proof. Recall that for any $x \in \mathcal{X}$ we have $\boldsymbol{g}_x^\mathsf{T}\boldsymbol{p} = \rho + \gamma p(x)$. Hence, our log-loss function $f(\boldsymbol{p}) = -\frac{1}{n}\sum_{x\in\mathcal{X}} n_x \log(\rho + \gamma p(x))$, whose gradient is the vector whose x-coordinate is $\frac{\partial f}{\partial p(x)} = \frac{-\gamma n_x}{\rho+\gamma p(x)}$. The Hessian of f is therefore the diagonal matrix whose diagonal entries are $\frac{\gamma^2 n_x}{(\rho+\gamma p(x))^2}$. Recall the definitions of γ and ρ: it is easy to see that $\gamma \geq \epsilon\rho$, and since $\epsilon < 1$ we also have that $e^\epsilon - 1 \leq 2\epsilon$, hence $\gamma \leq 2\epsilon \cdot \rho$. And so:

$$\begin{aligned}\nabla^2 f &\succeq \frac{\min_x\{n_x\}}{n} \cdot \frac{\gamma^2}{(\rho + 2\epsilon\rho\cdot 1)^2} I \\ &\succeq \min_x\{n_x\} \cdot \frac{\epsilon^2\rho^2}{\rho^2(1+2\epsilon)^2} I \succeq \min_x\{n_x\} \cdot \frac{\epsilon^2}{(1+2\epsilon)^2} I\end{aligned}$$

making f at least $(\frac{\epsilon^2}{9} \cdot \frac{\min_x\{n_x\}}{n})$-strongly convex. □

Corollary 20 (Corollary 5 restated.)**.** *In order to do identity testing under standard randomized response with confidence and power $\geq 2/3$, it is necessary and sufficient that we get $\Theta(\frac{T^{2.5}}{\epsilon^2\alpha^2})$ samples.*

Proof. For any $\boldsymbol{q} \in H_1$ it follows that $d_{\mathrm{TV}}(\varphi(\boldsymbol{p}), \varphi(\boldsymbol{q})) = \frac{1}{2}\|(\rho\mathbf{1}+\gamma\boldsymbol{p}) - (\rho\mathbf{1}+\gamma\boldsymbol{q})\|_1 = \frac{\gamma}{2}\|\boldsymbol{p}-\boldsymbol{q}\|_1 = \gamma \cdot d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \geq \gamma\alpha$. Recall that $\rho = \frac{1}{T-1+e^\epsilon}$ and $\gamma = \frac{e^\epsilon - 1}{T-1+e^\epsilon}$, and so, for $\epsilon < 1$ we have $\frac{1}{T+2\epsilon} \leq \rho \leq \frac{1}{T}$ and $\frac{\epsilon}{T+2} \leq \gamma \leq \frac{2\epsilon}{T}$, namely $\rho = \Theta(1/T)$ and $\gamma = \Theta(\epsilon/T)$. Next, we bound $\|\rho\mathbf{1} + \gamma\boldsymbol{p}\|_{\frac{2}{3}}$:

$$\begin{aligned}\left(\|\rho\mathbf{1} + \gamma\boldsymbol{p}\|_{\frac{2}{3}}\right)^{2/3} &= \sum_{x\in X} (\rho + \gamma p(x))^{\frac{2}{3}} \\ &\geq \sum_{x\in X} \max\{\rho^{2/3}, \gamma^{2/3} p(x)^{2/3}\} \\ &\geq \max\left\{T\rho^{2/3},\ \gamma^{2/3}\|\boldsymbol{p}\|_{\frac{2}{3}}^{2/3}\right\}\end{aligned}$$

Using the fact that $(a+b)^{2/3} \leq a^{2/3} + b^{2/3}$ (See Proposition 14 in Section A) we also get

$$\begin{aligned}\sum_{x\in X} (\rho + \gamma p(x))^{\frac{2}{3}} &\leq \sum_{x\in\mathcal{X}} \rho^{2/3} + \gamma^{2/3} p(x)^{2/3} \\ &= T\rho^{2/3} + \gamma^{2/3}\|\boldsymbol{p}\|_{\frac{2}{3}}^{2/3}\end{aligned}$$

It follows that the necessary and sufficient number of samples required for identity-testing under standard randomized response is proportional to

$$\begin{aligned}&\Theta\left(\frac{\|\rho\mathbf{1} + \gamma\boldsymbol{p}\|_{\frac{2}{3}}}{\gamma^2\alpha^2}\right) = \Theta\left(\left(\frac{\|\rho\mathbf{1} + \gamma\boldsymbol{p}\|_{\frac{2}{3}}^{2/3}}{\gamma^{4/3}\alpha^{4/3}}\right)^{3/2}\right) \\ &= \Theta\left(\left(\frac{T^{1/3} + \frac{\epsilon^{2/3}}{T^{2/3}}\|\boldsymbol{p}\|_{\frac{2}{3}}^{2/3}}{\frac{\epsilon^{4/3}}{T^{4/3}}\alpha^{4/3}}\right)^{\frac{3}{2}}\right) \overset{(*)}{=} \Theta\left(\frac{T^{2.5}}{\epsilon^2\alpha^2} + \frac{T\|\boldsymbol{p}\|_{\frac{2}{3}}}{\epsilon\alpha^2}\right)\end{aligned}$$

where the derivation marked by $(*)$ follows Proposition 15 in Section A. For any T-dimensional vector $\boldsymbol{x}$ with L_1-norm of 1 we have $\|x\|_{\frac{2}{3}} = \left(\sum_{i=1}^{T} x(i)^{\frac{2}{3}}\right)^{\frac{3}{2}} \leq \sqrt{T}$. Thus $\|p\| \leq \sqrt{T}$ and therefore the first of the two terms in the sum is the greater one. The required follows.

Comment: It is evident that the tester given by Valiant and Valiant (2014) solves (w.p. $\geq 2/3$) the problem of identity-testing in the randomized response model using $\Theta(T^{2.5}/\epsilon^2\alpha^2)$ samples. However, it is not a-priori clear why their lower bounds hold for our problem. After all, the set $\varphi(H_1)$ is only a subset of $\{\boldsymbol{q} : d_{\mathrm{TV}}(\varphi(\boldsymbol{p}), \boldsymbol{q}) \geq \gamma\alpha\}$. Nonetheless, delving into the lower bound of Valiant and Valiant, the collection of distributions which is hard to differentiate from $\boldsymbol{p}$ given $o\left(\|\boldsymbol{p}\|_{\frac{2}{3}}\alpha^2\right)$ samples is given by choosing suitable $\Delta(x)$ and then looking at the ensemble of distributions given by $\{p(x) \pm \Delta(x)\}$ for each $x \in \mathcal{X}$. Luckily, this ensemble is maintained under φ, mapping each such distribution to $\{\rho + \gamma p(x) \pm \gamma\Delta(x)\}$. The lower bound follows. □

Lemma 21 (Lemma 7 restated.)**.** *Suppose that n, the number of signals, is at least $\Omega(\frac{d^2}{\alpha^2\gamma^2}\max_j\{T^j\})$. Then the above procedure creates distributions $\boldsymbol{z}^j$ such that the*

product distribution $\bar{\boldsymbol{z}} = \boldsymbol{z}^1 \times \boldsymbol{z}^2 \times ... \times \boldsymbol{z}^d$ satisfies the following property. If the signals $y_1, ..., y_n$ were generated by $\varphi(\bar{\boldsymbol{p}})$ for some product-distribution $\bar{\boldsymbol{p}} = \boldsymbol{p}^1 \times ... \times \boldsymbol{p}^d$, then w.p. $\geq 8/9$ we have that $d_{\chi^2}(\varphi(\bar{\boldsymbol{z}}), \varphi(\bar{\boldsymbol{p}})) \leq \gamma^2\alpha^2/1000$.

Proof. Fix feature j. Let $\boldsymbol{p}^j$ be the marginal distribution of the distribution $\boldsymbol{p}$ which generated the samples (whether $\boldsymbol{p}$ belongs to H_0 or H_1) on the jth feature. It follows that projecting the signals onto their jth feature yields $y_1^j, y_2^j, ..., y_n^j$ which were generated using $\varphi(\boldsymbol{p}^j) = (1-\gamma)\boldsymbol{u}_{\mathcal{X}_j} + \gamma\boldsymbol{p}^j$. Kamath et al (2015) have shown that w.p. $\geq 8/9$ it holds that $d_{\chi^2}(\tilde{\boldsymbol{z}}^j, \varphi(\boldsymbol{p}^j)) \leq \frac{9(T^j-1)}{n+1}$. We now apply the linear transformation $\boldsymbol{z}^j = \frac{1}{\gamma}\left(I - \frac{1-\gamma}{T^j}1_{\mathcal{X}^j}\right)\tilde{\boldsymbol{z}}^j$ and similarly note that $\boldsymbol{p}^j = \frac{1}{\gamma}\left(I - \frac{1-\gamma}{T^j}1_{\mathcal{X}^j}\right)\varphi(\boldsymbol{p}^j)$. We note that $\boldsymbol{z}^j$ is a valid probability distribution: for each $x^j \in \mathcal{X}^j$ we have that $\tilde{z}^j(x^j) > \frac{1-\gamma}{T^j} + \gamma\tau$ hence $z^j(x^j) > \tau > 0$; and since $\sum_{x^j}\tilde{z}^j(x^j) = 1$ then $\sum_{x^j} z^j(x^j) = \frac{1}{\gamma}\sum_{x^j}\tilde{z}^j(x^j) - \frac{1-\gamma}{T^j} = \frac{1}{\gamma}(1-(1-\gamma)) = 1$. We thus bound the χ^2-divergence between $\boldsymbol{z}^j$ and $\boldsymbol{p}^j$:

$$d_{\chi^2}(\boldsymbol{z}^j, \boldsymbol{p}^j) = \sum_{x^j} \frac{\frac{1}{\gamma^2}\left(\tilde{z}^j(x^j) - \frac{1-\gamma}{T^j} - \varphi(p^j(x^j)) + \frac{1-\gamma}{T^j}\right)^2}{\frac{1}{\gamma}\left(\tilde{z}^j(x^j) - \frac{1-\gamma}{T^j}\right)}$$
$$= \frac{1}{\gamma}\sum_{x^j}\frac{(\tilde{z}^j(x^j) - \varphi(p^j(x^j)))^2}{\tilde{z}^j(x^j) - \frac{1-\gamma}{T^j}}$$
$$\overset{(*)}{\leq} \frac{30d}{\alpha\gamma^2}\sum_{x^j}\frac{(\tilde{z}^j(x^j) - \varphi(p^j(x^j)))^2}{\tilde{z}^j(x^j)} \leq \frac{270d}{\alpha\gamma^2}\cdot\frac{T^j-1}{n+1}$$

where the inequality in $(*)$ follows from the fact that $\tilde{z}^j(x^j) - \frac{1-\gamma}{T^j} > \gamma\tau = \frac{\alpha\gamma}{10d}\cdot\frac{1}{T^j}$ whereas $\tilde{z}^j(x^j) \leq \frac{1-\gamma}{T^j} + \gamma\cdot 1 \leq \frac{1+\gamma T^j}{T^j} \leq \frac{1+\gamma T}{T^j} \leq \frac{1+e^\epsilon-1}{T^j} \leq \frac{3}{T^j}$ as $\epsilon < 1$.

Next, we use the product lemma from (Reiss, 1989) (Lemma 3.3.10) (similar argument was made in (Acharya et al., 2015)). Assuming that $\sum_j d_{\chi^2}(\boldsymbol{z}^j, \boldsymbol{p}^j) \leq 1$ we now have that

$$d_{\chi^2}(\bar{\boldsymbol{z}}, \bar{\boldsymbol{p}}) \leq \exp\left(\sum_j d_{\chi^2}(\boldsymbol{z}^j, \boldsymbol{p}^j)\right) - 1 \leq 2\sum_j d_{\chi^2}(\boldsymbol{z}^j, \boldsymbol{p}^j)$$
$$\leq \frac{600d}{\alpha\gamma^2}\cdot\frac{\sum_j(T^j-1)}{n+1} \leq \frac{600d^2}{\alpha\gamma^2}\cdot\frac{\max_j\{T^j\}}{n+1}$$

Finally, we can obtain the bound we are after:

$$d_{\chi^2}(\varphi(\bar{\boldsymbol{z}}), \varphi(\bar{\boldsymbol{p}})) = \sum_{\bar{x}\in\mathcal{X}}\frac{(\rho + \gamma\bar{z}(\bar{x}) - \rho - \gamma\bar{p}(\bar{x}))^2}{\rho + \gamma\bar{z}(\bar{x})}$$
$$\leq \gamma^2\sum_{\bar{x}\in\mathcal{X}}\frac{\bar{z}(\bar{x}) - \bar{p}(\bar{x}))^2}{\gamma\bar{z}(\bar{x})} \leq \frac{600d^2}{\alpha\gamma}\cdot\frac{\max_j\{T^j\}}{n+1}$$

setting $n = \Omega(\frac{d^2}{\alpha^2\gamma^2}\max_j\{T^j\})$ gives the required bound of $d_{\chi^2}(\varphi(\bar{\boldsymbol{z}}), \varphi(\bar{\boldsymbol{p}})) \leq \alpha\gamma/1000$. □

Claim 22 (Claim 8 restated.)**.** *Assume the underlying distribution of the samples is $\boldsymbol{q}$ and that the number of signals is at least $n = \Omega(\frac{d^2(\max_j T^j)^2}{\alpha^2\gamma^2}\log(d\max_j T^j))$. Then w.p. $\geq 8/9$ our preprocessing step marks certain types each feature as "small" such that the probability (under $\boldsymbol{q}$) of sampling a type $(x^1, x^2, ..., x^d)$ such that $\exists j, x^j$ is small is $\leq \alpha/2$.*

Proof. Fix $\boldsymbol{q}$, the distribution that generated the samples. Thus, the signals were generated using the distribution $\varphi(\boldsymbol{q})$. Fix a feature j and look at the marginal of $\boldsymbol{q}$ with regards to the jth feature, $\boldsymbol{q}^j$. We call a type $x^j \in \mathcal{X}^j$ *infrequent* if $q^j(x^j) \leq \frac{\alpha}{3d}\cdot\frac{1}{T^j}$. We now argue that w.p. $\geq 8/9$ all types deemed as small by our preprocessing step (for all d features) are also infrequent. This follows immediately from the Hoeffding bound: If x^j is frequent then $\varphi(q^j(x^j)) = \frac{1-\gamma}{T^j} + \gamma q^j(x^j) \geq \frac{1-\gamma}{T^j} + \frac{\gamma\alpha}{3d}\cdot\frac{1}{T^j}$, but as x^j is deemed small the difference between $\frac{n_{x^j}}{n}$ and its expected value is at least $\frac{\alpha\gamma}{5d}\frac{1}{T^j}$, so Hoeffding assures us this event happens w.p. $\leq \exp(-2n\frac{\alpha^2\gamma^2}{25d^2}\frac{1}{(T^j)^2})$. Applying the union bound, the probability that any of $\sum_j T^j$ types that might be deemed as small is actually frequent is thus upper bounded by $\left(\sum_j T^j\right)\cdot\exp(-2n\frac{\alpha^2\gamma^2}{25d^2}\frac{1}{(T^j)^2})$. As $n = \Omega(\frac{d^2(\max_j T^j)^2}{\alpha^2\gamma^2}\log(d\max_j T^j))$ we infer that this bad event happens with probability $\leq 1/9$.

Now that we've established that all infrequent types are also deemed small in our pre-processing, we bound $\Pr_{\boldsymbol{q}}[(x^1, .., x^d) \ : \ \exists j \ s.t. \ x^j \text{ is small}]$ using the union bound:

$$\sum_j \Pr_{\boldsymbol{q}^j}[x^j \ : \ x^j \text{ is small}] \leq \sum_j \Pr_{\boldsymbol{q}^j}[x^j \ : \ x^j \text{ is infrequent}]$$
$$= \sum_j\sum_{x^j \text{ infrequent}}\Pr_{\boldsymbol{q}^j}[x^j] = \sum_j\sum_{x^j \text{ infrequent}}\frac{\alpha}{3d}\cdot\frac{1}{T^j}$$
$$\leq \sum_j\frac{\alpha}{3d} = \frac{\alpha}{3} \qquad \square$$

C. Missing Proofs: Non-Symmetric Scheme

Theorem 23 (9 restated.)**.** *For any convex set H, the problem of finding the max-likelihood $\boldsymbol{p} \in H$ generating the observed non-symmetric signals $(y_1, .., y_n)$ is poly-time solvable.*

Proof. Fix any $\boldsymbol{p} \in H$, a probability distribution on $\mathcal{X}$. Using the public G_i we infer a distribution on $\mathcal{S}$, as $\Pr[y_i = s] = \sum_{x\in\mathcal{X}}\Pr[y_i = s | y_i \text{ picked using } G_i\boldsymbol{e}_x]\cdot\Pr[\text{user } i \text{ has type } x] = \sum_{x\in\mathcal{X}}\boldsymbol{e}_s^\intercal G_i\boldsymbol{e}_x \cdot p(x) = \boldsymbol{e}_s^\intercal G_i\left(\sum_{x\in\mathcal{X}}p(x)\boldsymbol{e}_x\right) = \boldsymbol{e}_s^\intercal G_i\boldsymbol{p} \overset{\text{def}}{=} \boldsymbol{g}_i^{s\intercal}\boldsymbol{p}$, with $\boldsymbol{g}_i^s$ denoting the row of G_i corresponding to signal s.

Therefore, given the observed signals $(y_1, ..., y_n) \in \mathcal{S}^n$, the likelihood of any p is given by $L(\boldsymbol{p};\ y_1, ..., y_n) = \prod_i \boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p}$. Naturally, the function we minimize is the negation of the average log-likelihood, namely

$$f(\boldsymbol{p}) = -\frac{1}{n}\sum_i \log\left(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p}\right) \tag{1}$$

so the gradient of f is given by $\nabla f = -\frac{1}{n}\sum_i \frac{1}{\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p}}\boldsymbol{g}_i^{y_i}$, and thus, the Hessian of f is $\nabla^2 f = \frac{1}{n}\sum_i \frac{1}{(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p})^2}\boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}$. As the Hessian of f is a non-negative sum of rank-1 PSD matrices, we have that $\nabla^2 f$ is also a PSD, so f is convex. The feasibility of the problem $\min_{\boldsymbol{p}\in H} f(\boldsymbol{p})$ for a convex set H follows. □

Lemma 24 (Lemma 10 restated.)**.** *Fix $\delta > 0$ and assume that the number of signals we observe is $n = \Omega(T^3\log(1/\delta))$. Then w.p.$\geq 1-\delta$ it holds that the function $f(\boldsymbol{p})$ we optimize (as given in Equation* (1)*) is $\left(3\sqrt{T}\right)$-Lipfshitz and $\left(\frac{\eta^2}{2}\right)$-strongly convex over the subspace $\{\boldsymbol{x} :\ \boldsymbol{x}^\mathsf{T}\mathbf{1} = 0\}$ (all vectors orthogonal to the all-1 vector).*

Proof. Once the G_is have been picked, we view the n signals and are face with the maximum-likelihood problem as defined in Theorem 9. As a result of this particular construction, it is fairly evident to argue that the function f whose minimum we seek is Lipfshitz: the contribution of each user to the gradient of f is $(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p})^{-1}\boldsymbol{g}_i^{y_i}$. Since our optimization problem is over the probability simplex, then for each $\boldsymbol{p}$ we always have $\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p} \geq \frac{1}{2}-\eta > \frac{1}{4}$, whereas $\|\boldsymbol{g}_i^{y_i}\| \leq (\frac{1}{2}+\eta)\sqrt{T} \leq \frac{3}{4}\sqrt{T}$. Therefore, our function $f(\boldsymbol{p})$ is $(3\sqrt{T})$-Lipfshitz.

The argument which is hairier to make is that f is also $\Theta(\eta^2)$-strongly convex over the subspace orthogonal to the all-1 vector; namely, we aim to show that for any unit-length vector $\boldsymbol{u}$ such that $\boldsymbol{u}^\mathsf{T}\mathbf{1} = 0$ we have that $\boldsymbol{u}(\nabla^2 f)\boldsymbol{u} \geq \frac{\eta^2}{2}$. Since each coordinate of each $\boldsymbol{g}_i^s$ is non-negative and upper bounded by $\frac{1}{2}+\eta \leq 1$, then it is evident that for any probability distribution $\boldsymbol{p}$ and any unit-length vector $\boldsymbol{u}$ we have $\boldsymbol{u}^\mathsf{T}(\nabla^2 f(\boldsymbol{p}))\boldsymbol{u} \geq \frac{1}{n}\sum_i \frac{(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{u})^2}{(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{p})^2} \geq \frac{1}{n}\sum_i(\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{u})^2$, it suffices to show that the least eigenvalue of $\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)$ which is still orthogonal to $\mathbf{1}$ is at least $\eta^2/2$.

Let $h(\boldsymbol{v}_1, ..., \boldsymbol{v}_n)$ be the function that maps n vectors in $\{\frac{1}{2}+\eta, \frac{1}{2}-\eta\}^T$ to the least-eigenvalue of the matrix $\frac{1}{n}\sum_i \boldsymbol{v}_i\boldsymbol{v}_i^\mathsf{T}$ on $\mathcal{U} = \{\boldsymbol{x}\in\mathbb{R}^T :\ \boldsymbol{x}^\mathsf{T}\mathbf{1} = 0\}$. As ever, our goal is to argue that w.h.p we have that $h(\boldsymbol{g}_1^{y_1}, ..., \boldsymbol{g}_n^{y_n}) \approx \mathrm{E}_{\boldsymbol{g}_i^{y_i}}[h(\boldsymbol{g}_1^{y_1}, ..., \boldsymbol{g}_n^{y_n})]$. However, it is unclear what is $\mathrm{E}_{\boldsymbol{g}_i^{y_i}}[h(\boldsymbol{g}_1^{y_1}, ..., \boldsymbol{g}_n^{y_n})]$, and the reason for this difficulty lies in the fact that at each day i we pick either the "1"-signal or the "-1"-signal *based on the choice of $\boldsymbol{g}_i^1$ and $\boldsymbol{g}_i^{-1}$*. Namely, for each user i, the user's type x is chosen according to $\boldsymbol{p}$ and that is independent of G_i. However, once G_i is populated, the choice of the signal is determined by the column corresponding to type x. Had it been the case that each user's signal is fixed, or even independent of the entries of G_i, then it would be simple to argue that w.h.p. the value of h is $\geq \eta^2/3$. However, the dependence between that two row vectors we choose for users and the signal sent by the user makes arguing about the expected value more tricky.

So let us look at $\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}$. The key to unraveling the dependence between the vectors $\boldsymbol{g}_i^1, \boldsymbol{g}_i^{-1}$ and the signal y_i is by fixing the types of the n users in advance. After all, their type is chosen by $\boldsymbol{p}$ independently of the matrices G_i. Now, once we know user i is of type x_i then the signal y_i is solely a function of the x_i-column of G_i, *but the rest of the columns are independent of y_i*. Therefore, every coordinate $g_i^{y_i}(x')$ for any $x' \neq x_i$ is still distributed uniformly over $\{\frac{1}{2}+\eta, \frac{1}{2}-\eta\}$, and simple calculation shows that

$$\begin{aligned}
&\Pr[g_i^{y_i}(x_i) = \tfrac{1}{2}+\eta] \\
&= \sum_{s\in\{1,-1\}} \Pr[g_i^s(x_i) = \tfrac{1}{2}+\eta \text{ and } y_i = s] \\
&= \sum_{s\in\{1,-1\}} \Pr[y_i = s | g_i^s(x_i) = \tfrac{1}{2}+\eta]\Pr[g_i^s(x_i) = \tfrac{1}{2}+\eta] \\
&= 2\cdot\tfrac{1}{2}(\tfrac{1}{2}+\eta) = \tfrac{1}{2}+\eta
\end{aligned}$$

hence, $\Pr[g_i^{y_i}(x_i) = \frac{1}{2}-\eta] = \frac{1}{2}-\eta$ and so $\mathrm{E}[g_i^{y_i}(x_i)] = (\frac{1}{2}+\eta)^2 + (\frac{1}{2}-\eta)^2 = 2\frac{1}{4}+2\eta^2 = \frac{1}{2}+2\eta^2$. We thus have that $\mathrm{E}[\boldsymbol{g}_i^{y_i}] = \frac{1}{2}\mathbf{1} + 2\eta^2\boldsymbol{e}_{x_i}$.

Note that $\mathrm{E}[(g_i^{y_i}(x_i) - \frac{1}{2})^2] = \eta^2$ as we always have that $g_i^{y_i}(x_i) - \frac{1}{2} \in \{-\eta, \eta\}$. Thus, $\mathrm{E}[(\boldsymbol{g}_i^{y_i} - \frac{1}{2}\mathbf{1})(\boldsymbol{g}_i^{y_i} - \frac{1}{2}\mathbf{1})^\mathsf{T}] = \eta^2 I$. It follows that $\mathrm{E}[\boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}] = \eta^2 I - (\frac{1}{2})^2 1_{\mathcal{X}\times\mathcal{X}} + \frac{1}{2}\left(\mathrm{E}[\boldsymbol{g}_i^{y_i}]\mathbf{1}^\mathsf{T} + \mathbf{1}\mathrm{E}[\boldsymbol{g}_i^{y_i}]^\mathsf{T}\right) = (\frac{1}{2})^2 1_{\mathcal{X}\times\mathcal{X}} + \eta^2 I + \eta^2\left(\boldsymbol{e}_{x_i}\mathbf{1}^\mathsf{T} + \mathbf{1}\boldsymbol{e}_{x_i}^\mathsf{T}\right)$.

We therefore have that for any unit length $\boldsymbol{u}$ which is orthogonal to $\mathbf{1}$ we have $\mathrm{E}[\boldsymbol{u}^\mathsf{T}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\boldsymbol{u}] = \eta^2 n$, or in other words: $\mathrm{E}[P_{\mathcal{U}}(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}})] = \eta^2 I$ (with $P_{\mathcal{U}}$ denoting the projection onto the subspace $\mathcal{U}$). The concentration bound for any unit-length $\boldsymbol{u}\in\mathcal{U}$ follows from standard Hoeffding and Union bounds on a $\frac{1}{4}$-cover of the unit-sphere in $\mathcal{U}$. The argument is standard and we bring it here for completion.

Let $\boldsymbol{u}_1, ..., \boldsymbol{u}_m$ be a $\frac{1}{8}$-cover of the unit-sphere in $\mathcal{U}$. Standard arguments (see Vershynin (2010) Lemma 5.2) give that $m = O(20^T)$. Moreover, for any matrix M, suppose we know that for each $\boldsymbol{u}_j$ it holds that $\frac{3}{4}\eta^2 < \boldsymbol{u}_j^\mathsf{T}M\boldsymbol{u}_j \leq \|M\boldsymbol{u}_j\|$. Then let $\boldsymbol{u}$ be the unit-length $\boldsymbol{u}\in\mathcal{U}$ on which $\boldsymbol{u}^\mathsf{T}M\boldsymbol{u}$ is minimized (we denote the value at $\boldsymbol{u}$ as $\sigma_{\min}(M)$)

and let $\boldsymbol{u}_j$ its vector in the cover. Then we get

$$\begin{aligned}\sigma_{\min}(M) &= \boldsymbol{u}^\mathsf{T} M \boldsymbol{u} \\ &= \boldsymbol{u}_j^\mathsf{T} M \boldsymbol{u}_j - \boldsymbol{u}_j^\mathsf{T} M(\boldsymbol{u}_j - \boldsymbol{u}) + (\boldsymbol{u}_j - \boldsymbol{u})^\mathsf{T} M \boldsymbol{u} \\ &\geq \tfrac{3}{4}\eta^2 - \|M\boldsymbol{u}_j\| \cdot \tfrac{1}{8} - \tfrac{1}{8}\sigma_{\min}(M) \\ \Rightarrow \tfrac{9}{8}\sigma_{\min}(M) &\geq \tfrac{5}{8}\eta^2\end{aligned}$$

so $\sigma_{\min}(M) > \eta^2/2$. We therefore argue that for each $\boldsymbol{u}_j$ it holds that $\Pr[\boldsymbol{u}_j^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)\boldsymbol{u}_j < \frac{3}{4}\eta^2] < \delta/20^T$ and then by the union-bound the required will hold.

Well, as shown, $\mathrm{E}[\boldsymbol{u}_j^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)\boldsymbol{u}_j] = \eta^2$. Denote X_i as the random variable $(\boldsymbol{u}_j^\mathsf{T}\boldsymbol{g}_i^{y_i})^2$ and note that due to orthogonality to $\mathbf{1}$ we have that

$$0 \leq X_i = (\boldsymbol{u}_j^\mathsf{T}(\boldsymbol{g}_i^{y_i} - \tfrac{1}{2}\mathbf{1}))^2 \leq \|\boldsymbol{u}_j\|^2 \cdot \|\boldsymbol{g}_i^{y_i} - \tfrac{1}{2}\mathbf{1}\|^2 = \eta^2 T$$

The Hoeffding bound now assures us that $\Pr[\boldsymbol{u}_j^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)\boldsymbol{u}_j < \frac{3}{4}\eta^2] = \Pr[\boldsymbol{u}_j^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)\boldsymbol{u}_j - \mathrm{E}[\boldsymbol{u}_j^\mathsf{T}\left(\frac{1}{n}\sum_i \boldsymbol{g}_i^{y_i}\boldsymbol{g}_i^{y_i\mathsf{T}}\right)\boldsymbol{u}_j] < -\frac{1}{4}\eta^2] \leq \exp(\frac{-2n^2\eta^4/16}{n\cdot\eta^4 T^2}) = \exp(-n/8T^2) \leq \delta/20^T$ for $n = \Omega(T^3\ln(1/\delta))$. □

Proposition 25 (Proposition 12 restated.)**.**

$$\mathrm{E}[(\boldsymbol{\theta} - 2\eta\boldsymbol{f})(\boldsymbol{\theta} - 2\eta\boldsymbol{f})^\mathsf{T}] \preceq \tfrac{1}{n}I$$

Proof. The columns of each G_i are chosen independently, and moreover, the signal y_i depends only on a single column. Therefore, it is clear that for each $x \neq x'$ we have that $\mathrm{E}\left[(\theta(x) - 2\eta f(x))(\theta(x') - 2\eta f(x'))\right] = \mathrm{E}[\theta(x) - 2\eta f(x)]\cdot\mathrm{E}[\theta(x') - 2\eta f(x')] = 0$, so all the off-diagonal entries of the variance-matrix are 0. And for each type $x \in \mathcal{X}$ we have that

$$\begin{aligned}&\mathrm{E}[(\theta(x) - 2\eta f(x))^2] \\ &= \tfrac{1}{n^2}\sum_{i,i'} \mathrm{E}[\left(\tfrac{1}{\eta}(g_i^{y_i}(x) - \tfrac{1}{2}) - 2\eta f(x)\right)\left(\tfrac{1}{\eta}(g_{i'}^{y_{i'}}(x) - \tfrac{1}{2}) - 2\eta f(x)\right)] \\ &\overset{(*)}{=} \tfrac{1}{n^2}\sum_i \mathrm{E}[\left(\tfrac{1}{\eta}(g_i^{y_i}(x) - \tfrac{1}{2}) - 2\eta f(x)\right)^2] \\ &= \tfrac{1}{n^2}\sum_i \mathrm{E}[\left(\tfrac{1}{\eta}(g_i^{y_i}(x) - \tfrac{1}{2})\right)^2] - 4\eta f(x)\cdot\mathrm{E}[\tfrac{1}{\eta}(g_i^{y_i}(x) - \tfrac{1}{2})] + 4\eta^2 f(x)^2 \\ &\overset{(**)}{=} \tfrac{1}{n^2}\left(\sum_i 1 - 4\eta f(x)\cdot 2\eta\sum_i e_{x_i}(x) + \sum_i 4\eta^2 f(x)^2\right) \\ &= \tfrac{1}{n} - \tfrac{8}{n}\eta^2 f(x)^2 + \tfrac{4}{n}\eta^2 f(x)^2 = \tfrac{1-4\eta^2 f(x)^2}{n} \leq \tfrac{1}{n}\end{aligned}$$

where $(*)$ follows from the independence of the ith and the i'th sample, and $(**)$ holds because $\frac{1}{\eta}(g_i^{y_i}(x) - \frac{1}{2}) \in \{1, -1\}$. □

Theorem 26 (Theorem 13 restated.)**.** *Assume $\epsilon < 1$. Given $n = \Omega(\frac{T}{\alpha^2\epsilon^2}\left(T + d^2\sum_j T^j\right))$ iid drawn signals from the non-symmetric locally-private mechanism under a dataset whose types were drawn iid from some distribution $\boldsymbol{q}$, then w.p. $\geq 2/3$ over the matrices G_i we generate and the types in the dataset we have the following guarantee. If $\boldsymbol{q}$ is a product distribution, then $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) \leq \frac{\alpha}{2}$, and if $\boldsymbol{q}$ is α-far from any product distribution then $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) > \frac{\alpha}{2}$.*

Proof. The proof follows the derivations made at the proof of Theorem 11. For the time being, we assume the types of the n users are fixed and denote the frequency vector $\boldsymbol{f} = \langle\frac{n_x}{n}\rangle_T$. Moreover, for each feature j we denote the marginal frequency vector as $\boldsymbol{f}^j$. Recall that we have shown that $\mathrm{E}[\frac{1}{2\eta}\boldsymbol{\theta}] = \boldsymbol{f}$ and that $\mathrm{E}[(\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{f})(\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{f})^\mathsf{T}] \preceq \frac{1}{4\eta^2 n}I$.

Fix a feature j. The way we obtain $\boldsymbol{\theta}^j$ is by summing the entries of $\frac{1}{2\eta}\boldsymbol{\theta}$ for each type $x^j \in \mathcal{X}^j$. This can be viewed as a linear operator M^j, of dimension $T^j \times T$, where the x^j-row of M^j has 1 for each $x \in \mathcal{X}$ whose j-th feature is x^j and 0 anywhere else. Since each column has a single 1, it follows that for every two distinct types x^j and y^j, the dot-product of the x^j-row and the y^j-row of M^j is 0. Thus, since each row has exactly $\prod_{j'\neq j} T^{j'} = \frac{T}{T^j}$ ones, we have that $(M^j)(M^j)^\mathsf{T} = \frac{T}{T^j}I_{\mathcal{X}^j\times\mathcal{X}^j}$.

And so, for each feature j we have that $\mathrm{E}[\boldsymbol{\theta}^j] = \mathrm{E}[M^j(\frac{1}{\eta}\boldsymbol{\theta})] = M^j\boldsymbol{f} = \boldsymbol{f}^j$. Moreover, we also have $\mathrm{E}[(\boldsymbol{\theta}^j - \boldsymbol{f}^j)(\boldsymbol{\theta}^j - \boldsymbol{f}^j)] = \mathrm{E}[M^j(\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{f})(\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{f})^\mathsf{T}M^{j\mathsf{T}}] \preceq \frac{T}{4\eta^2 nT^j}I$. As a result, $\mathrm{E}[\|\boldsymbol{\theta}^j - \boldsymbol{f}^j\|^2] \leq \mathrm{trace}(\frac{T}{4\eta^2 nT^j}I) = \frac{T}{4\eta^2 n}$, and the Union-bound together with Chebyshev inequality gives that $\Pr[\exists\, j, \text{ s.t. } \|\boldsymbol{p}^j - \boldsymbol{f}^j\| > \frac{1}{2\eta}\sqrt{\frac{12dT}{n}}] < \sum_{j=1}^d \frac{1}{12d} = \frac{1}{12}$.

We now consider the randomness in $\boldsymbol{f}$. For every j we denote $\boldsymbol{q}^j$ as the marginal of $\boldsymbol{q}$ on the jth feature. Not surprisingly we have that $\mathrm{E}[\boldsymbol{f}] = \boldsymbol{q}$ and that for each feature $\mathrm{E}[\boldsymbol{f}^j] = M^j\mathrm{E}[\boldsymbol{f}] = \boldsymbol{q}^j$. Moreover, some calculations give that $\mathrm{E}[(\boldsymbol{f}^j - \boldsymbol{q}^j)(\boldsymbol{f}^j - \boldsymbol{q}^j)^\mathsf{T}] = \frac{1}{n}M^j\left(\mathrm{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\mathsf{T}\right)(M^j)^\mathsf{T} = \frac{1}{n}\left(\mathrm{diag}(\boldsymbol{q}^j) - (\boldsymbol{q}^j)(\boldsymbol{q}^j)^\mathsf{T}\right)$. As a result, for each j we have $\mathrm{E}[\|\boldsymbol{f}^j - \boldsymbol{q}^j\|^2] = \frac{1}{n}(1 - \|\boldsymbol{q}^j\|^2) \leq \frac{1}{n}$. Again, the union-bound and the Chebyshev inequality give that $\Pr[\exists j, \text{ s.t. } \|\boldsymbol{f}^j - \boldsymbol{q}^j\| > \sqrt{\frac{12dT}{n}}] < \frac{d}{12d} = \frac{1}{12}$.

And so, w.p. $\geq 5/6$ we get that for each features j we have $\|\boldsymbol{\theta}^j - \boldsymbol{q}^j\|_1 \leq \sqrt{T^j}\|\boldsymbol{\theta}^j - \boldsymbol{q}^j\| \leq \sqrt{T^j}(1 + \frac{1}{2\eta})\sqrt{\frac{12dT}{n}} \leq \sqrt{T^j}\cdot\sqrt{\frac{12dT}{\eta^2 n}}$, where in the last step we used the fact that $\eta < \frac{1}{2}$ hence $(1 + \frac{1}{2\eta}) < \frac{1}{\eta}$. We set n large enough to have $\|\boldsymbol{\theta}^j - \boldsymbol{q}^j\|_1 \leq 1$, and in particular it implies that for

any j we also have $\|\boldsymbol{\theta}^j\| \leq 2$. We thus apply the bound on the product of the $\boldsymbol{\theta}^j$s to derive that (Proposition 17 in the supplementary material) $\|\boldsymbol{\theta}^1 \times ... \times \boldsymbol{\theta}^d - \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d\|_1 \leq 2\sum_j \sqrt{T^j}\sqrt{\frac{12dT}{\eta^2 n}} \leq 2\sqrt{d} \cdot \sqrt{\sum_j T^j}\sqrt{\frac{12dT}{\eta^2 n}}$.

Moreover, in the proof of Theorem 11 we have shown that $\Pr[\|\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{q}\|_1 > \sqrt{\frac{12T^2}{\eta^2 n}}] < \frac{1}{6}$. In conclusion, setting $n = \Omega(\frac{T}{\alpha^2\eta^2}\left(T + d^2\sum_j T^j\right))$ we have that w.p. $\geq 2/3$ both of the following relations holds:

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d\|_1 \leq \tfrac{1}{2}\alpha$$
$$\|\tfrac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{q}\|_1 \leq \tfrac{1}{2}\alpha$$

Now, if $\boldsymbol{q}$ is a product distribution that we have that $\boldsymbol{q} = \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d$ and hence $\|\frac{1}{2\eta}\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1 \leq \alpha$. In contrast, if $\boldsymbol{q}$ is α-far (in total-variation distance, and so (2α)-far in L_1-norm) from any product distribution, then in particular $\|\boldsymbol{q} - \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d\|_1 \geq 2\alpha$ and we get that $\|\frac{1}{2\eta}\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1 \geq \|\boldsymbol{q} - \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d\|_1 - \|\bar{\boldsymbol{\theta}} - \boldsymbol{q}^1 \times ... \times \boldsymbol{q}^d\|_1 - \|\frac{1}{2\eta}\boldsymbol{\theta} - \boldsymbol{q}\|_1 \geq \alpha$. □

D. Additional Figures

For completion, we bring here the results of our experiments.

Figure 2 details the empirical distribution of $P(\boldsymbol{\theta})$ we get under the null-hypothesis, under different sample complexities ($n = \{10, 100, 1000, 10000\}$) for different sizes of domains ($T = \{10, 25, 50, 100\}$). Next to the curves we also draw the curve of the χ^T-distribution. Since all curves are essentially on top of one another, it illustrates our point: the distribution of $P(\boldsymbol{\theta})$ under the null-hypothesis is (very close) to the χ^2_T-distribution.

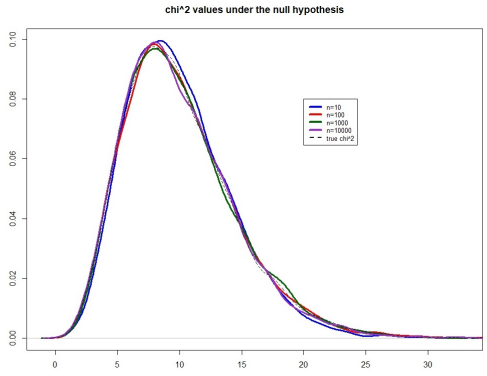
Figure 3 details the empirical distribution of $P(\boldsymbol{\theta})$ we get under the alternative-hypothesis, under different sample complexities ($n = \{2500, 5000, 7500, 10000, 200000\}$) for different TV-distances from the null-hypothesis ($\alpha = \{0.25, 0.2, 0.15, 0.1\}$). The results show the same pattern, as n increases, the distribution of $P(\boldsymbol{\theta})$ shifts away from the χ^2_T-distribution. This is clearly visible in the case where the total-variation distance is 0.25, and becomes less apparent as we move closer to the null-hypothesis.

Open Problems. The results of our experiment, together with the empirical results of the 3rd experiment (shown in Figure 1) give rise to the conjecture that the testers in Section 4.1 are not optimal. In particular, we suspect that the χ^2-based test we experiment with is indeed a valid tester of sample complexity $T^{1.5}/(\eta\alpha)^2$. Furthermore, there could be other testers of even better sample complexity. Both the improved upper-bound and finding a lower-bound are two important open problem for this setting. We suspect that the way to tackle this problem is similar to the approach of Acharya et al (2015); however following their approach is difficult for two reasons. First, one would technically need to give a bound on the χ^2-divergence between $\frac{1}{2\eta}\boldsymbol{\theta}$ and $\boldsymbol{q}$ (or $\boldsymbol{f}$). Secondly, and even more challenging, one would need to design a tester to determine whether the observed collection of random vectors in $\{1, -1\}^T$ is likely to come from the mechanism operating on a distribution close to $\frac{1}{2\eta}\boldsymbol{\theta}$. This distribution over vectors is a *mixture model* of product-distributions (but not a product distribution by itself); and while each product-distribution is known (essentially each of the T product distributions is a product of random $\{1, -1\}$ bits except for the x-coordinate which equals 1 w.p. $\frac{1}{2}+\eta$) it is the weights of the distributions that are either $\boldsymbol{p}$ or α-far from $\boldsymbol{p}$. Thus one route to derive an efficient tester can go through learning mixture models — and we suspect that is also a route for deriving lower bounds on the tester. A different route could be to follow the maximum-likelihood (or the loss-function f from Equation (1)), with improved convexity bounds proven directly on the L_1/L_∞-norms.

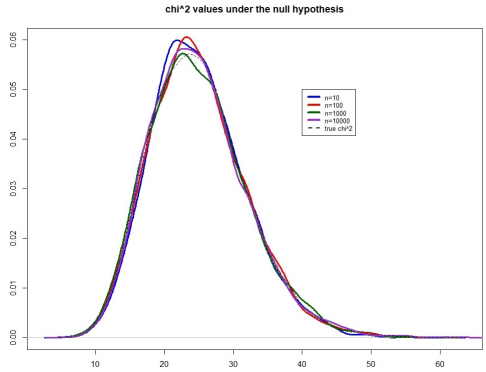
As explained in Section 4.2, we could not establish that

$$Q(\boldsymbol{\theta}) \stackrel{\text{def}}{=} n\sum_x \frac{(\frac{1}{2\eta}\theta(x) - \bar{\theta}(x))^2}{\bar{\theta}(x)}$$

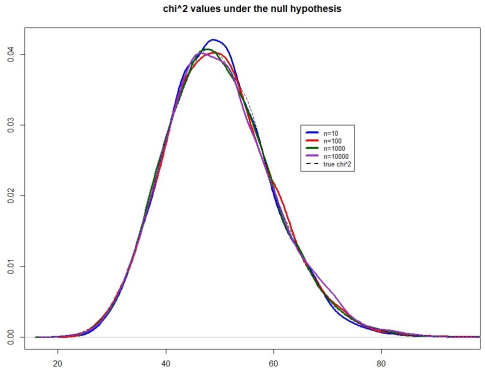
can serve as a test quantity, since we could not assess its asymptotic distribution. Nonetheless, we do believe it be a test quantity, as the following empirical results. We empirically measure the quantity $Q(\boldsymbol{\theta}) \stackrel{\text{def}}{=} n\sum_x \frac{(\frac{1}{2\eta}\theta(x) - \bar{\theta}(x))^2}{\bar{\theta}(x)}$ under the null ($\alpha = 0$) and the alternative ($\alpha = 0.25$) hypothesis with $n = 25,000$ samples in each experiment. The results under a variety of bin sizes are given in Figure 4. The results point to three facts: (1) the empirical distribution of Q under the null hypothesis is *not* a χ^2-distribution (it is not as centered around the mean and the tail is longer). (2) there is a noticeable gap between the distribution of $Q(\boldsymbol{\theta})$ under the null-hypothesis and under the alternative-hypothesis. Indeed, the gap becomes less and less clear under $25,000$ samples as the size of the domain increases, but it is present. (3) The empirical sample complexity required to differentiate between the null- and the alternative-hypothesis is quite large. Even for modest-size domains, $25,000$ samples weren't enough to create a substantial differentiation between the two scenarios. Designing a tester based on the quantity $Q(\boldsymbol{\theta})$ is thus left as an open problem.



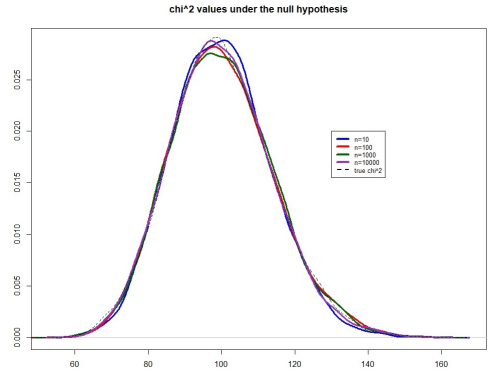
(a) #Types = 10



(b) #Types = 25

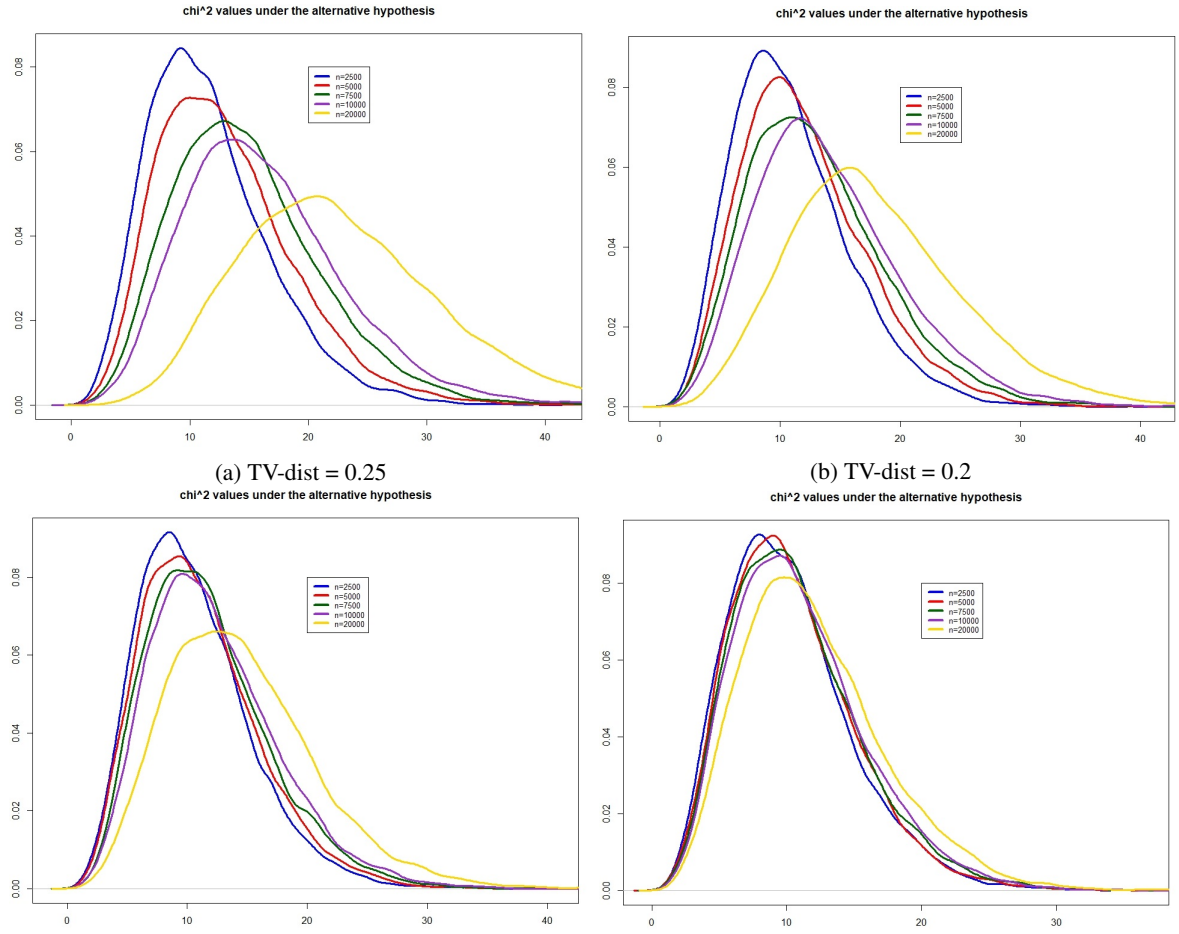


(c) #Types = 50



(d) #Types = 100

Figure 2: The empirical distribution of our test quantity under the null-hypothesis.
(Best seen in color) We ran our χ^2-based test under the null-hypothesis. Not surprisingly, the results we get seem to be taken from a χ^2-distribution (also plotted in a dotted black line). In all of the experiments we set $\epsilon = 0.25$.



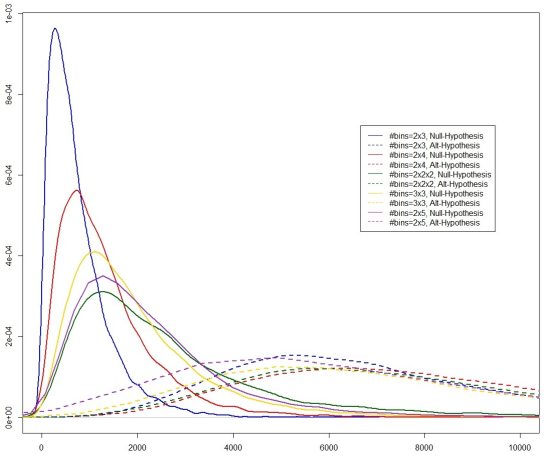
(a) TV-dist = 0.25

(b) TV-dist = 0.2

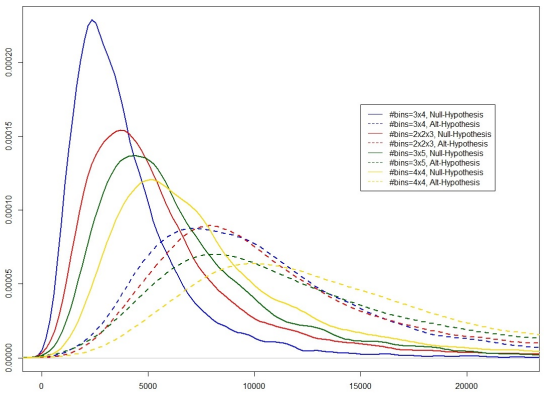
(c) TV-dist = 0.15

(d) TV-dist = 0.1

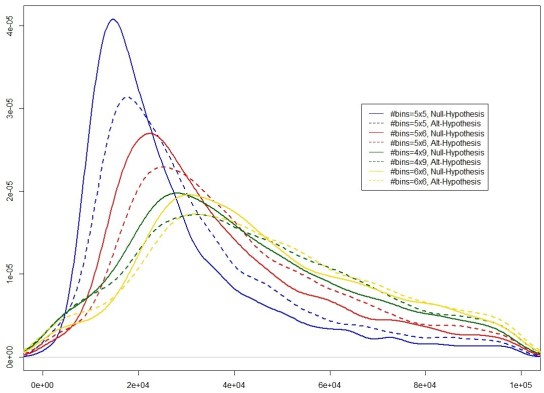
Figure 3: The empirical distribution of our test quantity under the alternative-hypothesis. (Best seen in color) We ran our χ^2-based test under the alternative-hypothesis with various choices of TV-distance. As the number of samples increases, the empirical distribution of the test-quantity becomes further away from the χ^2-distribution. In all of the experiments, the number of types is 10 and $\epsilon = 0.25$.



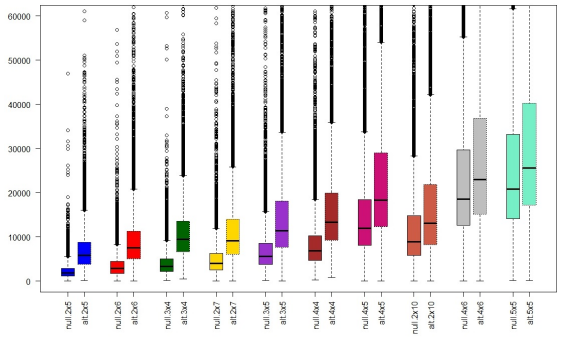
(a) Small Domain (between 6-10 types)



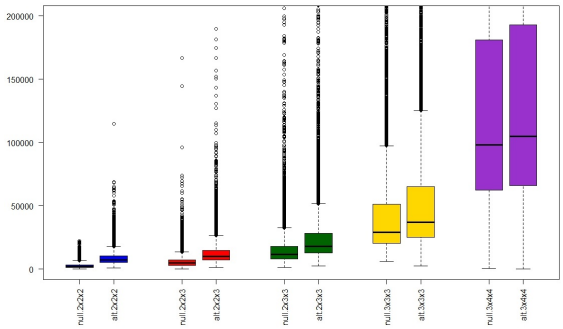
(b) Mid-Size Domain (12-16 types)



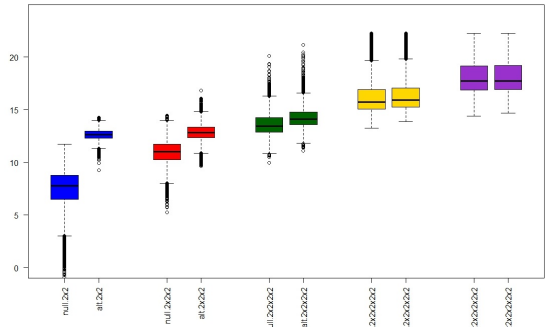
(c) Large Domain (25-36 types)



(d) 2-Dimensional Domains (12-30 types)



(e) 3-Dimensional Domains (8-48 types)



(f) Powers of 2 Size-Domains shown in log-scale (4-64 types)

Figure 4: The empirical distribution of $Q(\boldsymbol{\theta})$ for domains of multiple size, under the null- and the alternative-hypothesis.
(Best seen in color) We measured $Q(\boldsymbol{\theta})$ under both the null-hypothesis (solid line) and the alternative-hypothesis (dotted line) with various choices of domain sizes. As the size of the domain increases, it is evident the $25,000$ samples aren't enough to differentiate between the null and the alternative. In all of the experiments $\epsilon = 0.25$ and the alternative hypothesis is 0.25-far in TV-dist from a product distribution.