COLT2021 Tutorial – Recitation

Jayadev Acharya Clément Canonne Himanshu Tyagi

August 4, 2021

This document contains: (1) a short prelude, to check that some of the bounds obtained or claimed during the lecture part of the tutorial make sense; (2) a two-parter online component, where you will derive upper and lower bounds on some high-dimensional estimation problems under bandwidth or privacy constraints; (3) extra exercises, if you want additional material to go through at home. Detailed solutions will be made available, for we are not total monsters.

Exercise 0.1 (Sanity Checks)**.** *We have seen that, for discrete distribution estimation (in total variation/ℓ_1 distance) over $\{1, 2, \dots, d\}$, we could derive a lower bound of*

$$\Omega\left(\frac{d^2}{2^\ell \varepsilon^2}\right)$$

*on the sample complexity under ℓ-bits information constraints, and $\Omega\left(\frac{d^2}{\rho^2\varepsilon^2}\right), \Omega\left(\frac{d^2}{e^\rho\varepsilon^2}\right)$ under LDP constraints ($\rho \in (0, 1]$ and $\rho > 1$, respectively). Comment on those bounds: do they pass basic "sanity checks"? (*Hint: *Recall that,* without *constraints, the tight bound is known to be $\Theta\left(\frac{d}{\varepsilon^2}\right)$.)*

Solution. For $\ell = \log_2 d$, each user can actually send their full sample, since it only takes $\log_2 d$ bits to encode an element of $\{1, \dots, d\}$. So we want the lower bound $\Omega\left(d^2/(2^\ell\varepsilon^2)\right)$ to (1) decrease with ℓ (the bigger ℓ, the easier the task is, so the lower bound should become worse) and (2) retrieve the $\frac{d}{\varepsilon^2}$ *unconstrained* sample complexity when $\ell = \log_2 d$. Good, that's the case.

For the LDP constraint, this is similar: we retrieve the unconstrained sample complexity for $\rho = \log d$, and the lower bounds for $\rho > 1$ and $\rho \le 1$ are consistent for $\rho = 1$. Also, in all those cases the bound we obtain is higher with the local constraint than in the unconstrained case, which is… reassuring. □

1 Online: Mean Estimation, Upper and Lower Bounds

In what follows, we consider identity-covariance Gaussian distributions with bounded mean:

$$\mathcal{G}_d \coloneqq \{N(\mu, \mathbb{I}_d) : \mu \in \mathbb{R}^d, \|\mu\|_\infty \le 1\}$$

and Bernoulli mean products:

$$\mathcal{B}_d \coloneqq \{\otimes_{i=1}^d \operatorname{Rad}(\frac{1}{2}(\mu_i + 1)) : \mu \in \mathbb{R}^d, \|\mu\|_\infty \le 1\}$$

(product distributions over $\{\pm 1\}^d$ where the ith coordinate has mean μ_i (equals 1 with probability $\frac{\mu_i+1)}{2}$).

We will establish bounds on the sample complexity of mean estimation (under ℓ_2^2 loss) for both ℓ-bit communication constraints, and ρ-LDP. The first part of the recitation focuses on applying the framework discussed in the lectures to derive lower bounds (for the Bernoulli case) against interactive protocols. The second part will use some ideas evoked in the first part of the tutorial, as well as some general "tricks," to obtain matching upper bounds.

As a baseline to keep in mind, estimating the mean of either identity-covariance Gaussian distributions or Bernoulli products to ℓ_2 loss $\varepsilon \in (0,1]$ has sample complexity

$$\Theta\left(\frac{d}{\varepsilon^2}\right)$$

without any constraint.

1.1 Lower bound on Bernoulli Mean Estimation

We will use the general lower bound framework discussed in the tutorial to obtain a sample complexity lower bound for Bernoulli product mean estimation (i.e., the family $\mathcal{B}_d$) in ℓ_2^2 loss, under ℓ-bits communication and ρ-local privacy constraints. Here is what we want to use:

Consider a family of 2^d "hard distributions" $\mathcal{P} \subseteq \mathcal{B}_d$, indexed by $\mathcal{Z} := \{\pm 1\}^d$, which satisfies the assumptions. For us, $\mathcal{X} = \{\pm 1\}^d$ (our Bernoulli product distributions are over $\mathcal{X}$), and we will assume for notational simplicity that $\mathcal{Y}$ is discrete.

Assumption 1 (Additive loss)**.** *Fix $\varepsilon > 0$ and $\tau \in (0, 1/2]$. For all $z, z' \in \mathcal{Z}$,*

$$\ell_2(\mu_z, \mu_{z'}) \geq 4\varepsilon \left(\frac{\|z - z'\|_0}{\tau d}\right)^{1/2}$$

Assumption 2 (Densities exist)**.** *For all $z \in \mathcal{Z}$ and $i \in [d]$, there are $\alpha_{z,i} \in \mathbb{R}$ and $\phi_{z,i} \colon \mathcal{X} \to \mathbb{R}$ s.t.*

$$\frac{d\mathbf{p}_{z^{\oplus i}}}{d\mathbf{p}_z} = 1 + \alpha_{z,i}\phi_{z,i}$$

and $|\alpha_{z,i}| \leq \alpha$, where $\alpha \in \mathbb{R}$ is a fixed constant independent of z, i.

Assumption 3 (Bounded Ratios)**.** *There exists $\lambda \in [1, \infty]$ s.t.*

$$\sup_{z \in \mathcal{Z}} \sup_{y \in \mathcal{Y}} \sup_{W \in \mathcal{W}} \frac{\mathbb{E}_{X \sim \mathbf{p}_{z^{\oplus i}}}[W(y \mid X)]}{\mathbb{E}_{X \sim \mathbf{p}_z}[W(y \mid X)]} \leq \lambda$$

Assumption 4 (Orthonormality)**.** *For all $z \in \mathcal{Z}$ and $1 \leq i, j \leq d$,*

$$\mathbb{E}_{X \sim \mathbf{p}_z}[\phi_{z,i}(X)\phi_{z,j}(X)] = \begin{cases} 1 & \textit{if } i = j \\ 0 & \textit{otherwise.} \end{cases}$$

Assumption 5 (Subgaussianity)**.** *There is $\sigma \geq 0$ s.t., for all $z \in \mathcal{Z}$ and $1 \leq i, j \leq d$,*

$$\phi_z(X) := (\phi_{z,1}(X), \dots, \phi_{z,d}(X)) \in \mathbb{R}^d$$

is σ^2-subgaussian for $X \sim \mathbf{p}_z$, with independent coordinates.

Then we have the following:

Theorem 1. *Suppose Assumptions 1 to 3 hold, and let ε, τ as in Assumption 1. Suppose Π is an interactive protocol using $\mathcal{W}$ with n users, with expected ℓ_2^2 loss at most ε^2 over $\{\mathbf{p}_Z\}_Z$, when $Z \sim \mathrm{Rad}(\tau)^{\otimes d}$. Then, (1) if Assumption 4 holds,*

$$n \cdot \frac{\alpha^2}{d} \min\left(\lambda, \frac{1}{\tau}\right) \max_{z \in \mathcal{Z}} \max_{W \in \mathcal{W}} \sum_{y \in \mathcal{Y}} \frac{\mathrm{Var}_{X \sim \mathbf{p}_z}[W(y \mid X)]}{\mathbb{E}_{X \sim \mathbf{p}_z}[W(y \mid X)]} = \Omega(1)$$

and (2) if Assumption 5 holds,

$$n \cdot \frac{\alpha^2\sigma^2}{d} \min\left(\lambda, \frac{1}{\tau}\right) \max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}} H(\mathbf{p}_z^W) = \Omega(1)\,,$$

where $H(\mathbf{p}_z^W)$ is the Shannon entropy of the distribution $\mathbf{p}_z^W$ induced by $\mathbf{p}_z$ and W on $\mathcal{Y}$ (and thus is at most $\log_2|\mathcal{Y}|$).

Construction. Now, what should we choose as $\mathcal{P}$? *Discussion:* A natural choice is to have the discrepancy between means "spread out" uniformly over all coordinates (all coordinates of the mean to estimate are equally important) and to make each coordinate as unbiased as possible (since biased v. unbiased coin is the "hardest" in one dimension). This motivates setting, for some "suitable" $\gamma > 0$,

$$\mu_z := \gamma z \in [-1,1]^d, \qquad \mathbf{p}_z := \otimes_{i=1}^d \mathrm{Rad}(\frac{1}{2}(\mu_{z,i}+1)) \tag{1}$$

for every $z \in \{\pm 1\}^d$ (and $\mathcal{P} := \{\mathbf{p}_z\}_{z\in\{\pm1\}^d}$). Moreover, we'll consider the uniform prior of $\mathcal{Z}$, i.e., $Z \sim \mathrm{Rad}(1/2)^{\otimes d}$. *Discussion:* We don't have any sparsity constraint on μ, and all coordinates have a symmetric role, so uniform prior makes sense: each μ_i is chosen to be $\pm\gamma$ independently.

Exercise 1.1. *From Eq.* (1)*, write the expression for $\mathbf{p}_z(x)$, for arbitrary $z \in \{\pm1\}^d$ and $x \in \{\pm1\}^d$.*

Solution. Fix z. For $x \in \{\pm1\}^d$, we have, since $\gamma z_i = \mu_{z,i} = \mathbb{E}_{X\sim\mathbf{p}_z}[X_i] = \Pr_{\mathbf{p}_z}[X_i = 1] - \Pr_{\mathbf{p}_z}[X_i = -1]$ for all $1 \leq i \leq d$,

$$\begin{aligned}
\mathbf{p}_z(x) &= \prod_{i=1}^d \left(\Pr_{\mathbf{p}_z}[X_i = 1] \cdot \mathbf{1}_{\{x_i=1\}} + \Pr_{\mathbf{p}_z}[X_i = -1] \cdot \mathbf{1}_{\{x_i=-1\}} \right) \\
&= \prod_{i=1}^d \left(\frac{1+\mu_{z,i}}{2} \cdot \frac{1+x_i}{2} + \frac{1-\mu_{z,i}}{2} \cdot \frac{1-x_i}{2} \right) \\
&= \frac{1}{2^d} \prod_{i=1}^d (1+\gamma x_i z_i)
\end{aligned}$$

□

Exercise 1.2. *Given our choice of "dense" prior (uniform prior over $\mathcal{Z}$), which of the 5 assumptions do we not need to check?*

Solution. Since we take $\tau = 1/2$, we don't need to worry about Assumption 4: even $\lambda = \infty$ is OK for us, since our final bound when applying Theorem 1 will have $\min(\lambda, 1/\tau) \leq 1/\tau = 2$. *However,* you can check that in our case we *do* have Assumption 4 with $\lambda < \infty$: specifically, $\lambda = O(1)$. □

Exercise 1.3. *Given our construction of $\mathbf{p}_z$ (Eq.* (1)*) and τ, how should we set γ? (*Hint: *Assumption 1.)*

Solution. Since $\mu_z = \gamma z$, we have $\ell_2(\mu_z, \mu_{z'}) = \gamma\|z - z'\|_2 = 2\gamma\sqrt{\|z-z'\|_0}$, as $z, z' \in \{\pm1\}^d$. To satisfy Assumption 1, since $\tau = 1/2$ it suffices to set $\gamma \geq 2\sqrt{2}\varepsilon/\sqrt{d}$. Let's do that: $\gamma := 2\sqrt{2}\varepsilon/\sqrt{d}$. □

Exercise 1.4. *Show that if $\alpha_{z,i}, \phi_{z,i}$ satisfy Assumption 2, then $\mathbb{E}_{X\sim\mathbf{p}_z}[\phi_{z,i}(X)] = 0$.*

Solution. By Assumption 2,

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}_z}[\phi_{z,j}] = \frac{1}{\alpha_{z,i}}\mathbb{E}_{\mathbf{p}_z}\left[\frac{d\mathbf{p}_{z^{\oplus i}}}{d\mathbf{p}_z} - 1\right] &= \frac{1}{\alpha_{z,i}}\mathbb{E}_{\mathbf{p}_z}\left[\frac{d\mathbf{p}_{z^{\oplus i}}}{d\mathbf{p}_z} - 1\right] = \frac{1}{\alpha_{z,i}}\left(\mathbb{E}_{\mathbf{p}_z}\left[\frac{d\mathbf{p}_{z^{\oplus i}}}{d\mathbf{p}_z}\right] - 1\right) \\
&= \frac{1}{\alpha_{z,i}}\left(\mathbb{E}_{\mathbf{p}_{z^{\oplus i}}}[1] - 1\right) = 0
\end{aligned}$$

□

Exercise 1.5. *Find $\alpha_{z,i}$, $\phi_{z,i}$, α to satisfy Assumptions 2 and 4. (*Hint: *Start by finding the expression for the product $\alpha_{z,i}\phi_{z,i}$. Then compute $\mathbb{E}_{\mathbf{p}_z}\left[(\alpha_{z,i}\phi_{z,i})^2\right]$ and normalise to get $\alpha_{z,i}$, since $\mathbb{E}_{\mathbf{p}_z}\left[\phi_{z,i}^2\right]$ must equal 1 for Assumption 4. Finally, use the previous exercise to show Assumption 4 holds.)*

Solution. Fix z, i. From Exercise 1.1, we have for all $x \in \{\pm 1\}^d$

$$\frac{\mathbf{p}_{z^{\oplus i}}(x)}{\mathbf{p}_z(x)} = \frac{\prod_{j=1}^d (1+\gamma x_j z_j^{\oplus i})}{\prod_{j=1}^d (1+\gamma x_j z_j)} = \frac{1-\gamma x_i z_i}{1+\gamma x_i z_i} = 1 + \frac{-2\gamma x_i z_i}{1+\gamma x_i z_i} .$$

This tells us we should set $\alpha_{z,i}\phi_{z,i}(x) := \frac{-2\gamma x_i z_i}{1+\gamma x_i z_i}$. Now, since $z_i^2 = x_i^2 = 1$ for $z, x \in \{\pm 1\}^d$,

$$\begin{aligned}\mathbb{E}_{X\sim\mathbf{p}_z}\left[\left(\frac{-2\gamma X_i z_i}{1+\gamma X_i z_i}\right)^2\right] &= 4\gamma^2 \mathbb{E}_{X\sim\mathbf{p}_z}\left[\frac{1}{(1+\gamma X_i z_i)^2}\right] = 4\gamma^2 \mathbb{E}_{B\sim\mathrm{Rad}(\frac{1}{2}(1+\gamma z_i))}\left[\frac{1}{(1+\gamma B z_i)^2}\right] \\ &= 4\gamma^2\left(\frac{1+\gamma z_i}{2}\cdot\frac{1}{(1+\gamma z_i)^2} + \frac{1-\gamma z_i}{2}\cdot\frac{1}{(1-\gamma z_i)^2}\right) = \frac{4\gamma^2}{1-\gamma^2}\end{aligned}$$

the second inequality using the fact that $\mathbf{p}_z$ is a product distribution. This tells us that $\alpha_{z,i}$ should satisfy $\alpha_{z,i}^2 = \frac{4\gamma^2}{1-\gamma^2}$ (great! It's even independent of i), so we can set

$$\alpha = \alpha_{z,i} = \frac{2\gamma}{\sqrt{1-\gamma^2}}, \qquad \phi_{z,i}(x) := \frac{1}{\alpha}\cdot\frac{-2\gamma x_i z_i}{1+\gamma x_i z_i} = -\sqrt{1-\gamma^2}\cdot\frac{x_i z_i}{1+\gamma x_i z_i} = -\frac{x_i z_i - \gamma}{\sqrt{1-\gamma^2}} \tag{2}$$

We have so far Assumption 2, and that $\mathbb{E}_{\mathbf{p}_z}\left[\phi_{z,i}^2\right] = 1$ (our choice of $\alpha_{z,i} = \alpha$ guaranteed that). For Assumption 4, it only remains to show that $\mathbb{E}_{\mathbf{p}_z}[\phi_{z,i}\phi_{z,j}] = 0$ if $i \neq j$. But since $\mathbf{p}_z$ is a product distribution and $\phi_{z,i}$ only depends on x_i, we then have $\mathbb{E}_{\mathbf{p}_z}[\phi_{z,i}\phi_{z,j}] = \mathbb{E}_{\mathbf{p}_z}[\phi_{z,i}]\mathbb{E}_{\mathbf{p}_z}[\phi_{z,j}] = 0 \cdot 0 = 0$, the last equality by the previous exercise. □

We so far have shown Assumptions 1, 2 and 4 hold (and argued we didn't need Assumption 3). We already can apply part of Theorem 1!

Exercise 1.6. *Apply Theorem 1 to get the LDP lower bound for mean estimation to ℓ_2 loss ε: for $\rho \in (0,1]$, we must have*

$$n = \Omega\left(\frac{d^2}{\varepsilon^2\rho^2}\right) \tag{3}$$

and some *communication-constrained lower bound: for $\ell \geq 1$, we must have*

$$n = \Omega\left(\frac{d^2}{\varepsilon^2 2^\ell}\right) . \tag{4}$$

*(*Hint: *Use the statement from Exercise 2.4 for LDP, seen in the tutorial; for communication constraints, use* $\mathrm{Var}[W] \leq \mathbb{E}[W]$.*)*

Solution. The possibly tricky part is for Eq. (4). We use the fact that $0 \leq W(y \mid x) \leq 1$ for all $x \in \mathcal{X}, y \in \{0,1\}^\ell$ (it is, after all, a probability: the probability to send y upon seeing x) to write

$$\mathrm{Var}[W] \leq \mathbb{E}\left[W^2\right] \leq \mathbb{E}[W]$$

from which

$$\sum_{y\in\mathcal{Y}} \frac{\mathrm{Var}_{X\sim\mathbf{p}}[W(y \mid X)]}{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]} \leq \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]}{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]} = |\mathcal{Y}| = 2^\ell ,$$

and we can conclude from the first part of Theorem 1. □

The bound for communication constraints in Eq. (4) can be significantly improved, however. To do so, let's show Assumption 5 holds. First, some useful facts.

Recall that a mean-zero r.v. $U \in \mathbb{R}^d$ is σ^2-subgaussian if, for every unit vector $u \in \mathbb{R}^d$, the univariate r.v. $\langle u, U\rangle$ is σ^2-subgaussian: $\mathbb{E}\left[e^{t\langle u,U\rangle}\right] \le e^{\sigma^2 t^2/2}$ for all $t \in \mathbb{R}$. We also have the following:

Lemma 1 (Hoeffding's Lemma)**.** *Let X be a real-valued r.v. s.t. $a \le X \le b$ almost surely. Then*

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \le e^{\frac{t^2(b-a)^2}{8}}$$

for all $t \in \mathbb{R}$.

Exercise 1.7. *Show that Assumption 5 holds for $\sigma^2 = \frac{1+\gamma}{1-\gamma}$. Use then the second part of Theorem 1 the communication-constrained lower bound*

$$n = \Omega\left(\frac{d^2}{\varepsilon^2 \ell}\right) \tag{5}$$

for $\ell \ge 1$. (Hint: Use Hoeffding's lemma.)

Solution. Fix $z \in \{\pm 1\}^d$, and u s.t. $\|u\|_2^2 = 1$; and let $U := \phi_z(X)$ (where $X \sim \mathbf{p}_z$). By Exercise 1.4, the vector U has mean zero; by our expression for $\phi_{z,i}$ (which only depends on x_i) and the fact that $\mathbf{p}_z$ is a product distribution, the coordinates of $\phi_z(X)$ are independent. Finally, from Eq. (2) we see that

$$|\phi_{z,i}(X)| \le \frac{\sqrt{1-\gamma^2}}{1-\gamma} = \sqrt{\frac{1+\gamma}{1-\gamma}} =: \sigma$$

Thus we have, for $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{t\langle u,U\rangle}\right] &= \mathbb{E}_{X\sim\mathbf{p}_z}\left[e^{t\sum_{i=1}^d u_i \phi_{z,i}(X)}\right] = \mathbb{E}_{X\sim\mathbf{p}_z}\left[\prod_{i=1}^d e^{t u_i \phi_{z,i}(X)}\right] \\
&= \prod_{i=1}^d \mathbb{E}_{X\sim\mathbf{p}_z}\left[e^{t u_i \phi_{z,i}(X)}\right] && \text{(Independence of coordinates)} \\
&= \prod_{i=1}^d e^{\frac{t^2 u_i^2 \sigma^2}{2}} && \text{(Hoeffding's Lemma)} \\
&= e^{\frac{1}{2}t^2\sigma^2 \sum_{i=1}^d u_i^2} = e^{\frac{1}{2}t^2\sigma^2} && \text{(Since } \|u\|_2 = 1\text{)}
\end{aligned}$$

which shows that Assumption 5 is satisfied, as wanted. Applying the second part of Theorem 1 accordingly, noting that $\sigma^2 = O(1)$ and $\log_2|\mathcal{Y}| = \ell$ gives the result. □

In the next part, we will show that Eqs. (3) and (5) are optimal, and are actually achieved by *noninteractive* (even more, private-coin!) protocols.

1.2 Upper Bound on Gaussian and Bernoulli Mean Estimation

We now will complement the lower bounds (against interactive protocols) from the previous section by a *noninteractive, private-coin* protocol achieving the same sample complexity. The idea will be to first get the communication-constrained upper bound using 2 techniques:

- Reducing Gaussian to Bernoulli product mean estimation *(simpler! Everything is bits!)*
- Using SIMULATE-AND-INFER for Bernoulli products

and, finally, to use the communication-constrained algorithm for $\ell = 1$, along with a simple LDP protocol (Randomized Response) to get the LDP upper bound.

Exercise 1.8 (Reduction between Bernoulli Product and Gaussian)**.** *Suppose one has a mean estimation protocol for $\mathcal{B}_d$ under ℓ_2^2 loss with sample complexity n. Show that this implies a mean estimation protocol for $\mathcal{G}_d$ under ℓ_2^2 loss with sample complexity $O(n)$. Is the converse true? (*Hint: Erf *is well-behaved on* $[-1,1]$.*)*

Solution. Let $\mathbf{p} \in \mathcal{G}_d$ with mean $\mu(\mathbf{p}) = (\mu(\mathbf{p})_1, \dots, \mu(\mathbf{p})_d)$. For $X \sim \mathbf{p}$, let $Y = (\mathrm{sign}(X_i))_{i\in[d]} \in \{\pm 1\}^d$ be a random variable indicating the signs of the d coordinates of X. By independence of the coordinates of X, Y is distributed as a product Bernoulli distribution (in $\mathcal{B}_d$) with mean vector $\nu(\mathbf{p})$ given by

$$\nu(\mathbf{p})_i = 2 \Pr_{X\sim\mathbf{p}}[X_i > 0] - 1 = \mathrm{Erf}\left(\frac{\mu(\mathbf{p})_i}{\sqrt{2}}\right), \qquad i \in [d], \tag{6}$$

and, since $|\mu(\mathbf{p})_i| \le 1$, we have $\nu(\mathbf{p}) \in [-\eta,\eta]^d$, where $\eta := \mathrm{Erf}\big(1/\sqrt{2}\big) \approx 0.623$. Moreover, each user, given a sample from $\mathbf{p}$, can convert it to a sample from the corresponding product Bernoulli distribution. It remains to show that a good estimate for $\nu(\mathbf{p})$ gives a good estimate for $\mu(\mathbf{p})$.

Lemma 2. *Fix $\mathbf{p} \in \mathcal{G}_d$. For $\widehat{\nu} \in [-\eta,\eta]^d$, define $\widehat{\mu} \in [-1,1]^d$ by $\widehat{\mu}_i := \sqrt{2}\,\mathrm{Erf}^{-1}(\widehat{\nu}_i)$, for all $i \in [d]$. Then $\|\mu(\mathbf{p}) - \widehat{\mu}\|_2 \le \sqrt{\frac{e\pi}{2}} \cdot \|\nu(\mathbf{p}) - \widehat{\nu}\|_2$.*

Proof. By computing the maximum of its derivative,[1] we observe that the function Erf^{-1} is $\frac{\sqrt{e\pi}}{2}$-Lipschitz on $[-\eta,\eta]$. By the definition of $\widehat{\mu}$ and recalling Eq. (6), we then have

$$\|\mu(\mathbf{p}) - \widehat{\mu}\|_2^2 = \sum_{i=1}^d (\mu(\mathbf{p})_i - \widehat{\mu}_i)^2 = 2\sum_{i=1}^d \left(\mathrm{Erf}^{-1}(\nu_i) - \mathrm{Erf}^{-1}(\widehat{\nu}_i)\right)^2 \le \frac{e\pi}{2} \cdot \sum_{i=1}^d (\nu_i - \widehat{\nu}_i)^2,$$

where we used the fact that $\nu, \widehat{\nu} \in [-\eta,\eta]^d$. □

This concludes the argument. Note that the converse is false (or, at least, is unknown): at the very least, one can show that there is no conversion from Bernoulli to Gaussian which preserves distances between means. □

Exercise 1.9 (Upper Bound for Bernoulli Product)**.** *Use the exercise above to obtain a noninteractive, private-coin protocol for Gaussian mean estimation (under ℓ_2^2 loss) under ℓ-bit communication constraints with sample complexity $O\left(\frac{d^2}{\varepsilon^2 \min(\ell,d)}\right)$. (*Hint: *Simulate-and-Infer.)*

Solution. By the earlier exercise, it suffices to show the result for Bernoulli products. Great, now we have d bits to send (not real values), and ℓ we *can* send. Suppose $\ell \le d$. Since we have a product distribution $\mathbf{p}$ over $\{\pm 1\}^d$, we can use d/ℓ users to get *one* sample from $\mathbf{p}$, as follows: the first user sends the first ℓ bits of her sample, the second sends the coordinates $\ell+1, \dots, 2\ell$ of her own sample, etc. Concatenating those d/ℓ bits corresponding to disjoint groups of ℓ coordinates, we get a *bona fide* sample from $\mathbf{p}$.

We can then use the centralized (without constraints) algorithm for mean estimation of Bernoulli products, which takes $O\left(\frac{d}{\varepsilon^2}\right)$ samples. The overall number of users needed is $\frac{d}{\ell} \cdot O\left(\frac{d}{\varepsilon^2}\right) = O\left(\frac{d^2}{\varepsilon^2 \ell}\right)$. (This is, in effect, a version of Simulate-and-Infer for Bernoulli products, where the "simulate" part is quite simple.) Finally, if $\ell \ge d$, each user can send their full d-bit sample and we can use the centralized algorithm directly. □

Exercise 1.10 (From communication to local privacy)**.** *Deduce, for $\rho \in (0,1]$, the existence of a noninteractive, private-coin protocol for Gaussian mean estimation (under ℓ_2^2 loss) under ρ-LDP (no communication constraints) with sample complexity $O\left(\frac{d^2}{\varepsilon^2\rho^2}\right)$. (*Hint: *Case* $\ell = 1$, *and Randomized Response.)*

[1] Specifically, we have that $\max_{x\in[-\eta,\eta]} \mathrm{Erf}^{-1}(x) = 1/\sqrt{2}$ by definition of η and monotonicity of Erf. Recalling then that, for all $x \in [-\eta,\eta]$, $(\mathrm{Erf}^{-1})'(x) = \frac{1}{\mathrm{Erf}'(\mathrm{Erf}^{-1}(x))} = \frac{\sqrt{\pi}}{2} e^{(\mathrm{Erf}^{-1}(x))^2} \le \frac{\sqrt{\pi}}{2} e^{\frac{1}{2}}$, we get the Lipschitzness claim.

Solution. Recall that the *Randomized Response* (RR) mechanism (with parameter $\rho > 0$) is the channel $W\colon \{\pm 1\} \to \{\pm 1\}$ such that, for all $x, y \in \{\pm 1\}$,

$$W(y \mid x) = \begin{cases} \frac{e^\rho}{1+e^\rho} = \frac{1}{2} + \frac{\rho}{4} + o(\rho) & \text{if } x = y \\ \frac{1}{1+e^\rho} = \frac{1}{2} - \frac{\rho}{4} + o(\rho) & \text{if } x \neq y \end{cases}$$

This is ρ-LDP. Now, using the previous algorithm for $\ell = 1$, we can focus on learning the bias of each coordinate one by one (and see how many users we need to learn *one* coordinate of the mean μ to good accuracy, before multiplying that by d to get all coordinates). Since each message, after being sent through the RR channel, is basically a biased Bernoulli, with parameter $1/2 \pm O(\rho)$ depending on whether the true bit is ± 1, each "message bit" is a Bernoulli with parameter $1/2 + O(\rho\mu_i)$ and we need $O((\sqrt{d}/\rho\varepsilon)^2) = O(d/(\varepsilon^2\rho^2))$ users to get μ_i to an expected squared error (MSE) $(\varepsilon/\sqrt{d})^2 = \varepsilon^2/d$. Summing over all coordinates, since the MSE adds up, we need $d \cdot O(d/(\varepsilon^2\rho^2)) = O(d^2/(\varepsilon^2\rho^2))$ users to get μ to MSE $\mathbb{E}\left[\|\mu - \hat{\mu}\|^2\right] \leq \varepsilon^2$. (We can then convert that to a constant probability guarantee by using, e.g., Markov's inequality.) □

Exercise 1.11 (The boundedness assumption)**.** *Generalize the above exercises to*

$$\mathcal{G}_d(R) \coloneqq \{N(\mu, \mathbb{I}_d) : \mu \in \mathbb{R}^d, \|\mu\|_\infty \leq R\}$$

where $R > 0$ is a parameter given as input to the protocol. Can we set $R = \infty$?

Solution. The dependence on R will only show up in the reduction Gaussian $\rightsquigarrow$ Bernoulli: check the proof of Exercise 1.8 to see where. (Basically, the parameter $\nu = \operatorname{Erf}(1/\sqrt{2})$ governing the range of $\nu(\mathbf{p})$ will now depend mildly on R. But that part, by itself, is still alright! The "bad" dependence on R will then come from the change in the Lipschitzness constant of Erf^{-1}, which means that the constant $\sqrt{\frac{e\pi}{2}}$ in Lemma 2 will become an increasing function of R. Which means that, for large R, one will have to estimate $\nu(\mathbf{p})$ (the Bernoulli mean) to much higher accuracy to get the desired accuracy ε on $\mu(\mathbf{p})$.

However, having some dependence on R is necessary, so we cannot hope to do away with it and set $R = \infty$. Roughly speaking, one can use a packing argument to show that if the mean is assumed to have ℓ_∞ norm bounded by R, then the sample complexity (under either communication or local privacy constraints) must grow at least as $\log^{\Omega(1)} R$. □

2 Extra (offline): More exercises

The exercises marked with an asterisk do not have a solution provided. Feel free to email us if you're stuck!

2.1 Lower bounds: more sanity checks, and practise

Exercise 2.1. *Recall that we were able to derive a $\Omega\left(\frac{d^2}{\varepsilon^2 \ell}\right)$ lower bound for Gaussian mean estimation under ℓ-bits communication constraints. Why did we only get a 2^ℓ for discrete distribution estimation? (*Hint: *Check the subgaussianity+independence assumption. Also, the upper bound, for a "Well, duh!" answer.)*

Solution. The "Well, duh!" answer is that getting an ℓ dependence on the lower bound would contradict the *upper* bound of $O\left(\frac{d^2}{\varepsilon^2 2^\ell}\right)$, and also not match the known unconstrained setting bound (sample complexity of $\Theta\left(\frac{d}{\varepsilon^2}\right)$) when $\ell = \log_2 d$ (cf. Exercise 0.1).

A more precise answer is that we have neither independence, nor (nontrivial) subgaussianity with the lower bound construction for the discrete estimation case, where we have $|\phi_{z,i}| = O(\sqrt{d})$ (and corresponding, the best subgaussian parameter we can get is a "trivial" $\sigma^2 = \Theta(d)$). Note that the issue is not just in the independence part. Even under Poissonization, we would have had a 2^ℓ – the subgaussianity parameter is truly the issue. □

Exercise 2.2 $(*)$**.** *Suppose you are in a setting where each user sends their data via an erasure (oblivious) communication channel, which just "swallows" the message with fixed probability $\eta \in [0,1]$. Model this as a family $\mathcal{W}_\eta$ of local constraints, and derive a sample complexity lower bound for discrete distribution estimation (in total variation distance). (*Hint: $\mathcal{W}_\eta \subseteq \{W\colon [d] \to [d] \cup \{\bot\}\}$.*)*
Complement this with a matching upper bound.

2.2 Simulate-and-Infer

Exercise 2.3 (Simulate-and-Infer)**.** *Suppose you want to prove a (noninteractive) lower bound on the sample complexity of some estimation problem $\mathcal{P}$ under ℓ-bit communication constraints, of the form*

$$n = \Omega\left(\frac{f(\mathcal{P})}{2^\ell}\right).$$

Show that it suffices to prove it for $\ell = 1$. Does it extend to interactive protocols?

Solution. Suppose there exists a noninteractive protocol Π on n users under ℓ-bit communication constraints for $\mathcal{P}$. Then the output of user i (over the input $X \sim \mathbf{p}$) induces a distribution $\mathbf{p}'$ on $\{0,1\}^\ell$. Thus, by SIMULATE-AND-INFER, a sample from $\mathbf{p}'$ can be simulated by an (expected) $O(2^\ell)$ users with 1-bit communication constraints.

Overall, Π can be simulated using (in expectation) $N = O(2^\ell n)$ users with 1-bit communication constraints. This means that a lower bound $N = \Omega(f(\mathcal{P}))$ for 1-bit communication constraints implies the desired $n = \Omega(f(\mathcal{P})/2^\ell)$ for ℓ-bit constraints.
Yes, it extends. [Check it!] □

2.3 Connection to local privacy

Exercise 2.4. *Fix any $\rho > 0$, and suppose that $W\colon \mathcal{X} \to \mathcal{Y}$ is a ρ-locally private channel (for simplicity of notation, with discrete output space). Show that, for any probability distribution $\mathbf{p}$ on $\mathcal{X}$, we have*

$$\sum_{y\in\mathcal{Y}} \frac{\operatorname{Var}_{X\sim\mathbf{p}}[W(y \mid X)]}{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]} \leq \min\left((e^\rho - 1)^2, e^\rho\right).$$

(In particular, for $\rho \in [0,1]$ the RHS is $O(\rho^2)$.)

Solution. By definition of ρ-LDP, we have, for every y, x, x', $W(y \mid x) \leq e^\rho W(y \mid x')$ or, equivalently, $W(y \mid x) - W(y \mid x') \leq (e^\rho - 1)W(y \mid x')$. Taking expectations over $x' \sim \mathbf{p}$, we get

$$W(y \mid x) - \mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)] \leq (e^\rho - 1)\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]$$

We also have, similarly, $W(y \mid x) - W(y \mid x') \leq (1 - e^{-\rho})W(y \mid x) \leq (e^\rho - 1)W(y \mid x)$, so, taking expectations over $x \sim \mathbf{p}$, Taking expectations over $x' \sim \mathbf{p}$, we get

$$\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)] - W(y \mid x') \leq (e^\rho - 1)\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)].$$

Combining the two gives $|W(y \mid x) - \mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]| \leq (e^\rho - 1)\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]$. Squaring, and taking the expectation over $x \sim \mathbf{p}$ once more, we get $\operatorname{Var}_{X\sim\mathbf{p}}[W(y \mid X)] \leq (e^\rho - 1)^2 \mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]^2$. This holds for every $y \in \mathcal{Y}$; going back to our sum, we then have

$$\sum_{y\in\mathcal{Y}} \frac{\operatorname{Var}_{X\sim\mathbf{p}}[W(y \mid X)]}{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]} \leq (e^\rho - 1)^2 \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]^2}{\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]} = (e^\rho - 1)^2 \sum_{y\in\mathcal{Y}} \mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)] = (e^\rho - 1)^2$$

since $\sum_{y\in\mathcal{Y}} W(y \mid x) = 1$ for every $x \in \mathcal{X}$. This shows the first part of the statement; for the second, we bound (for a fixed $y \in \mathcal{Y}$) the variance as

$$\mathrm{Var}_{X\sim\mathbf{p}}[W(y \mid X)] \leq \mathbb{E}_{X\sim\mathbf{p}}\big[W(y \mid X)^2\big] \leq \sup_{x\in\mathcal{X}} W(y \mid x) \cdot \mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)] \leq e^{\rho}\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]^2$$

where the last inequality uses ρ-LDP to get $W(y \mid x) \leq e^{\rho}\mathbb{E}_{X\sim\mathbf{p}}[W(y \mid X)]$ for every x (as before, use the inequality for x, x' and take expectation over $x' \sim \mathbf{p}$). We conclude as earlier, summing over $y \in \mathcal{Y}$. □

2.4 Generalizing to per-user different LDP or comnunication constraints

This exercise asks you to generalize the arguments from Section 1 (both upper and lower bounds) to the setting where each user has a possibly different local privacy parameter $\rho_i \in (0, 1]$ (or a possibly different bandwidth constraint $\ell_i \geq 1$).

Exercise 2.5 $(*)$**.** *Suppose user i (for $1 \leq i \neq n$) has a local privacy parameter $\rho_i \in (0,1]$, and let $\bar{\rho}^2 := \frac{1}{n}\sum_{i=1}^n \rho_i^2$. Generalize the proofs of Section 1 for ℓ_2 mean estimation for Bernoulli products (or identity-covariance Gaussians) to obtain (i) a noninteractive, private-coin protocol with sample complexity*

$$O\left(\frac{d^2}{\varepsilon^2\bar{\rho}^2}\right)$$

and (ii) a matching sample complexity lower bound of $\Omega\left(\frac{d^2}{\varepsilon^2\bar{\rho}^2}\right)$ against interactive protocols. Prove the analogous statement if each user has a communication constraint ℓ_i, letting $\bar{\ell} := \frac{1}{n}\sum_{i=1}^n \ell_i$: i.e., establish a bound of $\Theta\left(\frac{d^2}{\varepsilon^2 \min(\ell, d)}\right)$.

(Hint: For the lower bound, recall (cf. tutorial, or check the paper!) that in Theorem 1 the "$n \sup_z \sup_{W\in\mathcal{W}}$" can be replaced by "$\sum_{i=1}^n \sup_z \sup_{W\in\mathcal{W}_i}$.")

2.5 From parameter estimation to estimating Z

Exercise 2.6 $(*)$**.** *Let $\mathcal{P}_\Theta := \{\mathbf{p}_\theta\}_{\theta\in\Theta}$ be a family of probability distributions over $\mathcal{X}$ parameterized by $\Theta \subseteq \mathbb{R}^k$, and $\{\mathbf{p}_z\}_{z\in\{\pm1\}^d} \subseteq \mathcal{P}_\Theta$. For simplicity, for each $z \in \{\pm1\}^d$ we denote by θ_z the parameter $\theta \in \Theta$ corresponding to $\mathbf{p}_z$. Suppose that there exists $\gamma > 0$ such that*

$$\|\theta_z - \theta_{z'}\|_2^2 \geq \gamma \cdot \|z - z'\|_0, \qquad \forall z, z' \in \{\pm1\}^d.$$

Discussion: *Compare this assumption to Assumption 1. Show that, given an estimator $\hat{\theta}\colon \mathcal{X}^n \to \Theta$ for $\mathcal{P}_\Theta$ such that*

$$\sup_{\theta\in\Theta} \mathbb{E}_{\mathbf{p}_\theta}\Big[\|\theta - \hat{\theta}\|_2^2\Big] \leq \varepsilon^2$$

one can obtain an estimator $\hat{Z}\colon \mathcal{X}^n \to \{\pm1\}^d$ such that

$$\sup_{z\in\{\pm1\}^d} \mathbb{E}_{\mathbf{p}_z}\Big[\|z - \hat{Z}\|_0\Big] \leq \frac{4\varepsilon^2}{\gamma}$$

What does this imply if $\{\mathbf{p}_z\}_{z\in\{\pm1\}^d}$ is defined by $\theta_z = \frac{10\varepsilon}{\sqrt{d}} z$? $\theta_z = \varepsilon z$?