Example 1. It’s bad to request too many unnecessary medical tests. You may have heard this advice before, and we can show why this is indeed sound advice. Suppose we run 15 (unnecessary) medical tests on a healthy patient, each corresponding to the detection of a certain disease. Suppose the tests are accurate so that the rate of false positives is only 2%, i.e., \( \Pr(\text{test } i \text{ positive } | \text{ no disease}) = 0.02 \). What is the chance that all 15 tests are negative given that the patient doesn’t have any disease?\[ \Pr(\text{all negative } | \text{ no diseases}) = 1 - (1 - 0.02)^{15} \approx 0.26. \]

In the following lecture, we investigate the problem of “exploration bias”, which arises in data analysis. That is, suppose we collect a large amount of data and perform many measurements/tests on it. Based on these measurements, we decide to report (seemingly) interesting/significant features of the data. In this scenario, not only does the value of the measurement depend on the data (evidently), but also the choice of which measurement to report depends on the data. This can bias our results, i.e., if we repeat the measurement on a fresh set of data, the results could differ significantly.

1 Definitions and Problem Statement

Let \( X \) denote the sample space, and \( D \in X^m \) denote the data set. Let \( \phi_i(D), i \in \{1, 2, \ldots, m\} \), denote hypothesis tests’ statistics, indexed by \( i \). (Since \( D \) is random, so is \( \phi_i(D) \).) The true mean is \( \mu_i = E[\phi_i(D)] \), where the expectation is over the randomness of the dataset. On a particular dataset \( D \), if \( T(D) = i \) is the selected test, the output of data exploration is the value \( \phi_{T(D)}(D) \). The reported value is thus \( E[\phi_{T(D)}(D)] \), resulting in a bias of \( E[\phi_{T(D)}(D)] - \mu_{T(D)} \).

Hence, we would like to bound
\[
|E[\phi_{T(D)}(D)] - \mu_{T(D)}|.
\]

Note that \( T \) is not necessarily a deterministic function of \( D \), rather there exists a conditional distribution \( P_{T|D} \). In the remainder, we will assume that \( T \) is chosen based on the measurements \( \phi = (\phi_1, \phi_2, \ldots, \phi_m) \) and suppress \( D \) in the notation. That is, we can rewrite (1) as
\[
|E[\phi_T - \mu_T]|.
\]

Example 2. Let \( \phi_1 \) and \( \phi_2 \sim \mathcal{N}(\mu, \sigma^2) \) i.i.d. Let \( T_0 = \arg\max_{i \in \{1, 2\}} \phi_i \) and generate \( T \) as follows: \( T = \begin{cases} T_0, & \text{with probability } 1 - p \\ 3 - T_0, & \text{with probability } p \end{cases} \) for some \( p \in [0, 1] \). Now to compute the
exploration bias, note that $E[\mu_T] = \mu$. On the other hand,

$$E[\phi_T] = \Pr(T = T_0)E[\phi_{T_0}] + \Pr(T = 3 - T_0)E[\phi_{(3-T_0)}]$$

$$= (1 - p)E[\max\{\phi_1, \phi_2\}] + pE[\min\{\phi_1, \phi_2\}].$$

Now let $S = \phi_1 + \phi_2$ and $\Delta = \phi_1 - \phi_2$. It is straightforward to check that $S \sim \mathcal{N}(2\mu, 2\sigma^2)$, $\Delta \sim \mathcal{N}(0, \sigma^2)$, max$\{\phi_1, \phi_2\} = S + |\Delta|$, and min$\{\phi_1, \phi_2\} = S - |\Delta|$. Then,

$$E[\phi_T] = \frac{1}{2} \left( (1 - p)E[S + |\Delta|] + pE[S - |\Delta|] \right)$$

$$= \frac{1}{2} \left( E[S] + (1 - 2p)E[|\Delta|] \right)$$

$$= \frac{1}{2} \left( 2\mu + (1 - 2p)\sqrt{\frac{4\sigma^2}{\pi}} \right)$$

$$= \mu + (1 - 2p)\sigma \sqrt{\frac{1}{\pi}}.$$

Hence, the exploration bias is given by

$$|E[\phi_T] - E[\mu_T]| = |1 - 2p|\sigma \sqrt{\frac{1}{\pi}}.$$

Note that for $p = \frac{1}{2}$, the bias is zero. Indeed, for $p = 1/2$, $T$ is independent of $(\phi_1, \phi_2)$; hence the index does not depend on the data, so we are not introducing any bias. As we decrease $p$ from $\frac{1}{2}$ to 0, we “increase the dependence” between $T_0$ and $(\phi_1, \phi_2)$, and the exploration bias increases accordingly.

As we saw in the above example, the exploration bias depends on the degree to which $T$ depends on $\phi$. Hence, we will use dependence measures to find good bounds on the bias.

We can rewrite (2) in a more abstract way. In particular, suppose we have an alphabet $Z$, two distributions on $Z$ denoted by $P$ and $Q$, and a function $f : Z \to R$. We want to bound

$$|E_P[f(Z)] - E_Q[f(Z)]| \quad (3)$$

In the above setup $Z = R^m \times \{1, 2, \ldots, m\}$ (where $R^m$ and $\{1, 2, \ldots, m\}$ represent the sets in which $\phi$ and $T$ live, respectively), $f(\phi, t) = \phi_t$, $P = P_{\phi T}$ (i.e., the joint distribution of $T$ and $\phi$), and $Q = P_\phi P_T$ (i.e., the product of the marginals of $T$ and $\phi$).

## 2 $L_1$-Distance Bound

We have already seen a result somewhat similar to a bound on (3). In particular,
Lemma 1. Let \( P \) and \( Q \) be two probability mass functions on a finite set \( Z \). Then,

\[
\|P - Q\|_1 = 2 \max_{S \subseteq Z} P(S) - Q(S).
\]

Note that for any subset \( S \), \( P(S) \) can also be seen as \( E_P[f_S(Z)] \) where \( f_S(z) = \begin{cases} 1, & z \in S, \\ 0, & z \notin S \end{cases} \).

Moreover, the proof of Lemma 1 can be simply modified to show the following:

Lemma 2.

\[
\|P - Q\|_1 = 2 \max_{f:Z \to [0,1]} E_P[f(Z)] - E_Q[f(Z)].
\]

Proof: Let \( A = \{z \in Z : P(z) \geq Q(z)\} \). Then,

\[
E_P[f(Z)] - E_Q[f(Z)] = \sum_{z \in A} f(z) (P(z) - Q(z)) + \sum_{z \notin A} f(z) (P(z) - Q(z)) \\
\leq \sum_{z \in A} (P(z) - Q(z)) \\
= \frac{\|P - Q\|_1}{2}.
\]

Equality can be achieved if we choose \( f(z) = \begin{cases} 1, & z \in A, \\ 0, & z \notin A \end{cases} \). ■

Remark. The form of the equality in Lemma 2 is called the variational representation of \( L_1 \)-distance. More generally, we can represent any convex function (such as the \( L_1 \)-distance) as the supremum of affine functions.

We now have a bound on (3) for bounded \( f \):

Corollary 3. For any distributions \( P \) and \( Q \) of a finite set \( Z \), and any function \( f:Z \to [0,1] \), we have

\[
|E_P[f(Z)] - E_Q[f(Z)]| \leq \frac{\|P - Q\|_1}{2}.
\]

Exercise 1. The statement of Lemma 2 does not include the absolute value. Verify that the corollary follows by applying Lemma 2 twice: once using \( f \), and another using \( g = 1 - f \).

As noted earlier, our initial setup corresponds to choosing \( P \) to be some joint distribution \( P_{XY} \), and \( Q \) to be the product of the marginals \( P_X P_Y \). Then, the closer \( P_{XY} \) is to \( P_X P_Y \), the closer they are to independence (i.e., the less \( Y \) depends on \( X \)), which makes the exploration bias smaller, as captured in the corollary.

One disadvantage of the above bound is that it is restricted to bounded functions. And as noted in the remark, the main property that allowed us to derive such bound is the convexity of the \( L_1 \)-distance. Hence, we can derive similar bounds using other convex dependence measures. In particular, we will turn to the KL divergence.
3 Mutual Information Bound

The following lemma is called the Donsker-Varadhan variational representation of KL divergence.

**Lemma 4.** Let \( p \) and \( q \) be two probability density functions. Then,

\[
D(p||q) = \sup_{f:R \to R} \left\{ E_p[f(Z)] - \log E_q \left[ e^{f(Z)} \right] \right\}
\]

**Remark.** \( \log \) is taken to the base \( e \) (both in the computation of \( D(p||q) \) and in the right-hand side).

**Proof:**
1) Suppose \( D(p||q) < +\infty \) and consider any function \( f \). Then,

\[
E_p[f(Z)] - \log E_q \left[ e^{f(Z)} \right] = E_p \left[ \log e^{f(Z)} \right] - \log E_q \left[ e^{f(Z)} \right]
\]

\[
= E_p \left[ \log \frac{e^{f(Z)}}{E_q \left[ e^{f(Z)} \right]} \right]
\]

\[
= E_p \left[ \log \left( \frac{e^{f(Z)}}{E_q \left[ e^{f(Z)} \right]} \right) \right]
\]

\[
= D(p||q) + E_p \left[ \log \left( \frac{q e^{f(Z)}}{p E_q \left[ e^{f(Z)} \right]} \right) \right]
\]

\[
= D(p||q) - E_p \left[ \log \frac{p e^{f(Z)}}{q E_q \left[ e^{f(Z)} \right]} \right]
\]

Now note that \( \int q \frac{e^{f(Z)}}{E_q \left[ e^{f(Z)} \right]} dz = \frac{1}{E_q \left[ e^{f(Z)} \right]} \int q e^{f(Z)} dz = 1 \). Hence, the second term is also a KL divergence. Then,

\[
E_p[f(Z)] - \log E_q \left[ e^{f(Z)} \right] = D(p||q) - D(p||\tilde{q}) \leq D(p||q).
\]

Equality is achieved if we set \( f = \log \frac{p}{q} \).

2) Suppose \( D(p||q) = +\infty \). Then, we need to show that the supremum is also \( +\infty \). \( D(p||q) = +\infty \) implies that there exists a set \( A \) such that \( p(A) > 0 \) and \( q(A) = 0 \). Choose \( f = \lambda 1_{\{z \in A\}} \). Then, \( E_p[f(Z)] - \log E_q \left[ e^{f(Z)} \right] = \lambda p(A) \). Taking \( \lambda \to +\infty \) yields the result.

Before introducing the main bound, we need to introduce the concept of subgaussian distributions.

**Definition 1.** \( Z \) is called \( \sigma^2 \)-subgaussian if \( \log E \left[ e^{\lambda Z} \right] \leq \frac{\lambda^2 \sigma^2}{2} \) for all \( \lambda \in R \).

**Lemma 5.** If \( Z \) is \( \sigma^2 \)-subgaussian, then \( E[Z] = 0 \) and \( E[Z^2] \leq \sigma^2 \).

**Proof:** Pick \( \lambda \) such that \( \lambda E[Z] > 0 \). Using the Taylor expansion of the exponential, we get

\[
1 + \frac{1}{2} \lambda^2 \sigma^2 + O(\lambda^4) \geq e^{\lambda^2 \sigma^2/2} \geq E \left[ e^{\lambda Z} \right] \geq 1 + \lambda E[Z] + \frac{1}{2} \lambda^2 E[Z^2] + O(\lambda^3).
\]

The result follows by taking \( \lambda \to 0 \).
Exercise 2. For $Z \sim \mathcal{N}(0, \sigma^2)$, show that $E[e^{\lambda Z}] = e^{\lambda^2 \sigma^2/2}$. Hence, $Z$ is $\sigma^2$-subgaussian.

We are now ready to prove the main bound for this lecture on the exploration bias $|E[\phi_T - \mu_T]|$, where $\phi = (\phi_1, \phi_2, \ldots, \phi_m)$ and $T \in \{1, 2, \ldots, m\}$.

**Theorem 6.** Suppose for each $i \in \{1, 2, \ldots, m\}$, $\phi_i - \mu_i$ is $\sigma^2$-subgaussian. Then,

$$|E[\phi_T - \mu_T]| \leq \sigma \sqrt{2I(T; \phi)}.$$  

**Remark.** As expected, if $T$ is independent of $\phi$, then the exploration bias is zero. If $T$ does not depend “too much” on $\phi$, as captured by mutual information, then we can guarantee a small bias.

**Proof:** Fix $T = i$. We will use Lemma 4 with the distributions $P_{\phi_i|T=i}$ (the distribution of $\phi_i$ conditioned on the choice of $T$ being $i$) and $P_{\phi_i}$ (the prior distribution of $\phi_i$). For some $\lambda \in \mathbb{R}$, let $f = \lambda (\phi_i - \mu_i)$. Then, it follows from Lemma 4 that

$$D(P_{\phi_i|T=i} \| P_{\phi_i}) \geq \lambda \left( E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i \right) - \log E_{P_{\phi_i}} \left[ e^{\lambda (\phi_i - \mu_i)} \right]$$

$$\geq \lambda \left( E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i \right) - \lambda^2 \sigma^2/2,$$

where the second inequality follows from the $\sigma^2$-subgaussianity assumption. Since $\lambda$ was arbitrary, we get

$$D(P_{\phi_i|T=i} \| P_{\phi_i}) \geq \sup_{\lambda \in \mathbb{R}} \left\{ \lambda \left( E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i \right) - \lambda^2 \sigma^2/2 \right\}$$

$$= \frac{(E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i)^2}{2\sigma^2}.$$  

Hence,

$$|E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i| \leq \sigma \sqrt{2D(P_{\phi_i|T=i} \| P_{\phi_i}).}$$

Finally,

$$|E[\phi_T - \mu_T]| = \sum_{i=1}^{m} \Pr(T = i) |E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i|$$

$$\leq \sum_{i=1}^{m} \Pr(T = i) |E_{P_{\phi_i|T=i}} [\phi_i] - \mu_i|$$

$$\leq \sum_{i=1}^{m} \Pr(T = i) \sigma \sqrt{2D(P_{\phi_i|T=i} \| P_{\phi_i})}$$

$$\leq \sum_{i=1}^{m} \Pr(T = i) \sigma \sqrt{2D(P_{\phi|T=i} \| P_{\phi})}$$

$$\leq \sigma \sqrt{2 \sum_{i=1}^{m} \Pr(T = i) D(P_{\phi|T=i} \| P_{\phi})}$$

$$= \sigma \sqrt{2I(P_{\phi|T} \| P_{\phi})} = \sigma \sqrt{2I(T; \phi)},$$

where (a) and (c) follow from Jensen’s inequality, and (b) follows from the data processing inequality. 


Exercise 3. Show that, if \( \phi_i - \mu_i \) is \( \sigma_i^2 \)-subgaussian for each \( i \in \{1, 2, \ldots, m\} \), then

\[
|E[\phi_T - \mu_T]| \leq \sqrt{E[\sigma_T^2]} \sqrt{2I(T; \phi)}.
\]

Let’s revisit the initial example:

Example 3. Let \( \phi_1 \) and \( \phi_2 \) \( \sim \mathcal{N}(\mu, \sigma^2) \) i.i.d. Let \( T_0 = \arg\max_{i \in \{1, 2\}} \phi_i \) and generate \( T \) as follows: \( T = \begin{cases} T_0, & \text{with probability } 1 - p \\ 3 - T_0, & \text{with probability } p \end{cases} \) for some \( p \in [0, 1] \).

Since \( \phi_i - \mu_i \sim \mathcal{N}(0, \sigma^2) \), it is \( \sigma^2 \)-subgaussian, thus satisfying the assumption of Theorem 6.

To compute \( I(T; \phi) \):

\[
I(T; \phi) = H(T) - H(T|\phi).
\]

Since \( \phi_1 \) and \( \phi_2 \) are i.i.d, then \( \Pr(T_0 = 1) = \Pr(T_0 = 2) = \frac{1}{2} \). Hence, \( H(T) = \log 2 \).

Since both \( \phi - T_0 - T \) and \( T_0 - \phi - T \) are Markov chains, we get \( H(T|\phi, T_0) = H(T|\phi) \) and \( H(T|\phi, T_0) = H(T|T_0) \). Hence,

\[
I(T; \phi) = H(T) - H(T|\phi) = \log 2 - H(T|T_0) = \log 2 - H(p).
\]

Hence, by the above theorem,

\[
|E[\phi_T - \mu_T]| \leq \sigma \sqrt{2(\log 2 - H(p))}.
\]

Example 4. Suppose \( \phi_i \sim \mathcal{N}(0, \sigma^2) \) i.i.d. for \( i \in \{1, 2, \ldots, m\} \), and \( T = \arg\max_i \phi_i \). Then,

\[
I(T; \phi) = H(T) = \log m,
\]

and

\[
E \left[ \max \{\phi_1, \phi_2, \ldots, \phi_m\} \right] \leq \sigma \sqrt{2 \log m}.
\]