Probabilistic Method: Hoeffding's Inequality and Differential Privacy
Lecturer: Hubert Chan
Date: 27 May 2012

1 Hoeffding's Inequality

1.1 Approximate Counting by Random Sampling

Suppose there is a bag containing red balls and blue balls. You would like to estimate the fraction of red balls in the bag. However, you are only allowed to sample randomly from the bag with replacement. The straightforward method is to make T random samples and use the fraction of red balls in the random samples as an estimate of the true fraction. We are interested in estimation with certain additive error ϵ, i.e., if the true fraction is p, it is enough to return a number in the range $[p-\epsilon, p+\epsilon]$.

The question is: how many samples are enough so that with probability at least $1-\delta$, our estimate is within an additive error of $\epsilon > 0$ from the true value?

Let $Y := \sum_{t \in [T]} Z_t$, where Z_t is the $\{0,1\}$-random variable that takes value 1 if the t-th sample is red. The estimator is $\frac{Y}{T}$. Suppose the true fraction of red balls in the bag is p. Then, we want to find how large T needs to be such that the following is satisfied:

$$Pr[|\frac{Y}{T} - p| \geq \epsilon] \leq \delta.$$

Theorem 1.1 (Hoeffding's Inequality) *Suppose $X_1, X_2, \ldots, X_n$ are independent real-valued random variables, such that for each i, X_i takes values from the interval $[a_i, b_i]$. Let $Y := \sum_i X_i$. Then, for all $\alpha > 0$,*

$$Pr[|Y - E[Y]| \geq n\alpha] \leq 2\exp(-\frac{2n^2\alpha^2}{\sum_i R_i^2}),$$

where $R_i := b_i - a_i$.

Using the Hoeffding's Inequality, with $X_i \in [0,1]$ and $E[X_i] = p$, we have

$Pr[|\frac{1}{T}\sum_i X_i - p| \geq \alpha] \leq 2\exp(-2T\alpha^2)$.

Hence, in order to estimate the fraction of red balls with additive error at most α and failure probability at most δ, it suffices to use $T = \Theta(\frac{1}{\alpha^2}\log\frac{1}{\delta})$.

1.2 Proof of Hoeffding's Inequality

We use the technique of moment generating function to prove the Hoeffding's Inequality. For simplicity, we prove a slightly weaker result:

$Pr[|Y - E[Y]| \geq n\alpha] \leq 2\exp(-\frac{n^2\alpha^2}{2\sum_i R_i^2})$.

Here are the three steps.

1.2.1 Transform the Inequality into a Convenient Form

We use the inequality $Pr[|Y - E[Y]| \geq n\alpha] \leq Pr[Y - E[Y] \geq n\alpha] + Pr[Y - E[Y] \leq -n\alpha]$, and give bounds for the two probabilities separately. Here, we consider $Pr[Y - E[Y] \geq n\alpha]$. The other case is similar.

Observe that the expectation of Y does not appear in the upperbound. It would be more convenient to first translate the variable X_i. Define $Z_i := X_i - E[X_i]$. Observe that $E[Z_i] = 0$. Moreover, since both X_i and $E[X_i]$ are in the interval $[a_i, b_i]$, it follows that Z_i takes values in the interval $[-R_i, R_i]$.

For $t > 0$, we have

$Pr[Y - E[Y] \geq n\alpha] = Pr[t \sum_i Z_i \geq tn\alpha]$.

1.2.2 Using Moment Generating Function and Independence

Applying the exponentiation function to both sides of the inequality, we follow the standard calculation, using independence of the Z_i's.

$$\begin{aligned} Pr[t\sum_i Z_i \geq tn\alpha] &= Pr[\exp(t\sum_i Z_i) \geq \exp(tn\alpha)] \\ &\leq \exp(-tn\alpha) \cdot E[\exp(t\sum_i Z_i)] \\ &\leq \exp(-tn\alpha) \cdot \prod_i E[\exp(tZ_i)] \end{aligned}$$

The next step is the most technical part of the proof. Recall that we want to find some appropriate function $g_i(t)$ such that $E[\exp(tZ_i)] \leq \exp(g_i(t))$. All we know about Z_i is that $E[Z_i] = 0$ and Z_i takes value in $[-R_i, R_i]$.

Think of a simple (but non-trivial) random variable that has mean 0 and takes values in $[-R_i, R_i]$. Consider $\widehat{Z}_i$ that takes value R_i with probability $\frac{1}{2}$ and $-R_i$ with probability $\frac{1}{2}$. Then, it follows that $E[\exp(t\widehat{Z}_i)] = \frac{1}{2}(e^{tR_i} + e^{-tR_i}) \leq \exp(\frac{1}{2}t^2R_i^2)$, the last inequality follows from the fact that for all real x, $\frac{1}{2}(e^x + e^{-x}) \leq e^{\frac{1}{2}x^2}$.

Therefore, for such a simple $\widehat{Z}_i$, we have a nice bound $E[\exp(t\widehat{Z}_i)] \leq \exp(g_i(t))$, where $g_i(t) := \frac{1}{2}t^2R_i^2$.

Using Convexity to Show that the Extreme Points are the Worst Case Scenario. Intuitively, $\widehat{Z}_i$ is the worst case scenario. Since we wish to show measure concentration, it is a bad case if there is a lot of variation for Z_i. However, we have the requirement that $E[Z_i] = 0$ and $Z_i \in [-R_i, R_i]$. Hence, intuitively, Z_i has the most variation if it takes values only at the extreme points, each with probability $\frac{1}{2}$. We formalize this intuition using the convexity of the exponentiation function.

Definition 1.2 *A function $f : \mathbb{R} \to \mathbb{R}$ is convex if for all $x, y \in R$ and $0 \leq \lambda \leq 1$,*

$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$.

Fact 1.3 *If a function is doubly differentiable and $f''(x) \geq 0$, then f is convex.*

We use the fact that the exponentiation function $x \mapsto e^{tx}$ is convex. First, we need to express Z_i as a convex combination of the end points, i.e., we want to find $0 \leq \lambda \leq 1$ such that $Z_i = \lambda R_i + (1-\lambda)(-R_i)$. We have $\lambda := \frac{Z_i + R_i}{2R_i}$.

Using the convexity of the function $x \mapsto e^{tx}$, we have

$$\begin{aligned}
\exp(tZ_i) &= \exp(t(\lambda R_i + (1-\lambda)(-R_i))) \\
&\leq \lambda \exp(tR_i) + (1-\lambda)\exp(-tR_i) \\
&= (\frac{1}{2} + \frac{Z_i}{2R_i})\exp(tR_i) + (\frac{1}{2} - \frac{Z_i}{2R_i})\exp(-tR_i).
\end{aligned}$$

Take expectation on both sides, and observing that $E[Z_i] = 0$, we have

$E[\exp(tZ_i)] \leq \frac{1}{2}(\exp(tR_i) + \exp(-tR_i)) = E[\exp(t\widehat{Z}_i)]$, as required.

1.2.3 Picking the Best Value for t

We have shown that for all $t > 0$,

$Pr[Y - E[Y] \geq n\alpha] \leq \exp(\frac{1}{2}t^2 \sum_i R_i^2 - n\alpha t)$, and we want to find a value of t that minimizes the right hand side.

Note that the exponent is a quadratic function in t and hence is minimized when $t := \frac{n\alpha}{\sum_i R_i^2} > 0$. Hence, in this case, we have

$Pr[Y - E[Y] \geq n\alpha] \leq \exp(-\frac{n^2\alpha^2}{2\sum_i R_i^2})$, as required.

1.3 Exercise

Estimating the (Unknown) Fraction of Red Balls. Suppose a bag contains an unknown number of red balls (assume there is at least one red ball) and you are only allowed to sample (with replacement) uniformly at random from the bag. Design an algorithm that, given $0 < \epsilon, \delta < 1$, with failure probability at most δ, returns an estimate of the fraction of red balls with **multiplicative** error at most ϵ, i.e., if the real fraction is p, the algorithm returns a number $\widehat{p}$ such that $|\widehat{p} - p| \leq \epsilon p$. Give the number of random samples used by your algorithm.

2 Differential Privacy

2.1 Motivation

A *statistical database* is a database used for statistical analysis. For example, a database containing information about graduates of a certain university can answer questions like: what is the average salary of the graduates? Statistical database are widely used due to the enormous social value they provide: the previously mentioned database benefits the society in helping students to choose whether to go to that university, or how funding should be distributed among universities. However, the statistics released might cause leakage of sensitive information. Therefore, a big challenge is to maintain individual privacy, while providing useful aggregate statistical information about a certain group.

In 1977 the statistician Tore Dalenius gave a privacy goal for statistical databases: anything that can be learned about a member in the statistical database, should also can be learned without access to the database. However, as the following example illustrates, as long as the statistical database is useful, the goal is not achievable if the adversary has auxiliary information (information not obtained from the statistical database). Suppose we know that some student's X salary is $10,000 more than the average, we can know his/her salary by querying about the average salary. Note that in this case, even if X does not join the database, we can still approximately know the average salary and hence know his/her sensitive information.

Notation. Let $\mathcal{U}$ be the set of possible user data points. A database of n users contains the data points for each user and can be viewed as a vector in $\mathcal{U}^n$. We use $\mathcal{D} := \mathcal{U}^n$ to denote the collection of all possible databases with n users. We consider whether releasing the output of some function $f : \mathcal{D} \to \mathcal{O}$ will compromise an individual's privacy.

Example. Suppose there are n users, each of which holds a private bit from $\mathcal{U} = \{0, 1\}$. Given a database $X \in \mathcal{U}^n$, suppose the function of interest is $\mathsf{sum}(X) = \sum_{i=1}^n X_i$. At first sight, it might seem that releasing the sum does not violate an individual's privacy, because the sum does not directly reveal any individual's private bit. However, in reality, users' data can be in many databases. Suppose users 1 to $n - 1$ also participate in another database which also releases the sum. Then, combining the results of the two sums, the private bit of user n can be accurately calculated!

Differential privacy defines privacy in a different sense: to minimize the increased risk of the sensitive information leakage due to one's joining in the statistical database. A private mechanism that answers queries to statistical databases can achieve this goal by introducing randomness such that when two databases differ by only one single user, the output produced have similar distributions. Note that a differentially private mechanism encourages individuals to participate in statistical databases, and thus enhancing the social benefit provided by them.

2.2 Formal Definition

Let $\mathcal{U}$ be the set of possible data points and $\mathcal{D} := \mathcal{U}^n$ be the collection of databases with n users. Two databases $X^{(1)} \in \mathcal{D}$ and $X^{(2)} \in \mathcal{D}$ are called *neighboring* (denoted by $X^{(1)} \sim X^{(2)}$), if they differ by at most one coordinate.

Definition 2.1 (ϵ-differential privacy) *A randomized mechanism (function) $\mathcal{M} : \mathcal{D} \to \mathcal{O}$ preserves ϵ-differential privacy, if for any two neighboring databases $X^{(1)} \sim X^{(2)}$, and any possible set of output $\mathcal{S} \subseteq \mathcal{O}$, the following hold:*

$$\Pr[\mathcal{M}(X^{(1)}) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(X^{(2)}) \in \mathcal{S}],$$

where the randomness comes from the coin flips of $\mathcal{M}$.

Remark 2.2 We observe the following.

1. Since $\mathcal{M}$ is a randomized mechanism, $\mathcal{M}(X)$ is a distribution of outputs in $\mathcal{O}$.
2. Interchanging the roles of $X^{(1)}$ and $X^{(2)}$, we also have:
   $\Pr[\mathcal{M}(X^{(2)}) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(X^{(1)}) \in \mathcal{S}]$.
3. If $\mathcal{O}$ is a countable set, then we can also have the inequality for each $x \in \mathcal{O}$,
   $\Pr[\mathcal{M}(X^{(1)}) = x] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(X^{(2)}) = x]$.
4. The inequality means that the distributions $\mathcal{M}(X^{(1)})$ and $\mathcal{M}(X^{(2)})$ are close, and hence by observing the output, it is not possible to tell for certain whether the output comes from database $X^{(1)}$ or $X^{(2)}$.

2.3 Achieving Differential Privacy

Given a set $\mathcal{D}$ of databases, a deterministic function $f : \mathcal{D} \to \mathbb{R}^d$, and a privacy parameter ϵ, we want to find an ϵ-differentially private version of f, i.e., to find a (random) function $\widehat{f} : \mathcal{D} \to \mathbb{R}^d$ that has the following properties:

1. **Privacy.** The function $\widehat{f}$ preserves ϵ-differential privacy,
2. **Utility.** For any database $X \in \mathcal{D}$, with high probability, $\widehat{f}(X)$ is close to $f(X)$.

Observe that the trivial function $\widehat{f} \equiv 0$ satisfies ϵ-differential privacy. However, this will not be very useful. Here is one way to define utility.

Definition 2.3 ((λ, δ)-useful) *Let $0 < \delta < 1$ and $\lambda > 0$ be utility parameters. Let $f : \mathcal{D} \to \mathbb{R}^d$ be a deterministic function and $\widehat{f} : \mathcal{D} \to \mathbb{R}^d$ be a randomized function. We say $\widehat{f}$ is (λ, δ)-useful with respect to f, if for any database $X \in \mathcal{D}$, with probability at least $1 - \delta$, for each $i \in [d]$, $|f_i(X) - \widehat{f}_i(X)| \leq \lambda$.*

2.4 Achieving differential privacy by adding Laplace noise

One way to convert a function $f : \mathcal{D} \to \mathbb{R}^d$ into a differentially private version is to add independent random noise to each of its coordinates. Intuitively, the more $f(X)$ changes when we change one coordinate of X, the larger the random noise is needed to hide the difference. We use ℓ_1-sensitivity to formally measure the maximum difference between the values of f of any two neighboring databases.

Definition 2.4 (ℓ_1-Sensitivity) *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a deterministic function. The ℓ_1-sensitivity of f, denoted by Δf, is*

$$\max_{X^{(1)} \sim X^{(2)}} ||f(X^{(1)}) - f(X^{(2)})||_1 = \max_{X^{(1)} \sim X^{(2)}} \sum_{i=1}^{d} |f_i(X^{(1)}) - f_i(X^{(2)})|$$

We use random variables sampled from Laplace distributions as the random noise.

Definition 2.5 (Laplace Distribution) *Let $b > 0$. We denote by $\mathsf{Lap}(b)$ the Laplace distribution such that the probability density function at z is $\frac{1}{2b}\exp(-\frac{|z|}{b})$.*

Laplace distribution has the following properties.

Theorem 2.6 *Let $b > 0$ and let γ be a random variable sampled from $\mathsf{Lap}(b)$. Then,*

1. $E[\gamma] = 0$,
2. $var[\gamma] = 2b^2$,
3. *for any* $\lambda > 0$, $\Pr[|\gamma| > \lambda] = \exp(-\frac{\lambda}{b})$.

Proof:

1.

$$\begin{aligned}E[\gamma] &= \int_{-\infty}^{\infty} \frac{x}{2b}\exp\left(-\frac{|x|}{b}\right)dx = \int_{-\infty}^{0} \frac{x}{2b}\exp\left(\frac{x}{b}\right)dx + \int_{0}^{\infty} \frac{x}{2b}\exp\left(\frac{-x}{b}\right)dx\\ &= \int_{\infty}^{0} \frac{-x}{2b}\exp\left(\frac{-x}{b}\right)d(-x) + \int_{0}^{\infty} \frac{x}{2b}\exp\left(\frac{-x}{b}\right)dx\\ &= -\int_{0}^{\infty} \frac{x}{2b}\exp\left(\frac{-x}{b}\right)dx + \int_{0}^{\infty} \frac{x}{2b}\exp\left(\frac{-x}{b}\right)dx = 0\end{aligned}$$

2.

$$\begin{aligned}E[\gamma^2] &= \int_{-\infty}^{\infty} \frac{x^2}{2b}\exp\left(-\frac{|x|}{b}\right)dx = 2\int_{0}^{\infty} \frac{x^2}{2b}\exp\left(-\frac{|x|}{b}\right)dx\\ &= -\int_{0}^{\infty} x^2\exp\left(-\frac{x}{b}\right)d\left(-\frac{x}{b}\right) = -(x^2\exp\left(-\frac{x}{b}\right)|_0^\infty - \int_0^\infty 2x\exp\left(-\frac{x}{b}\right)dx)\\ &= 2\int_0^\infty x\exp\left(-\frac{x}{b}\right)dx = -2b\int_0^\infty x\exp\left(-\frac{x}{b}\right)d\left(-\frac{x}{b}\right)\\ &= -2b\left(x\exp\left(-\frac{x}{b}\right)|_0^\infty - \int_0^\infty \exp\left(-\frac{x}{b}\right)dx\right) = 2b\int_0^\infty \exp\left(-\frac{x}{b}\right)dx\\ &= -2b^2\int_0^\infty \exp\left(-\frac{x}{b}\right)d\left(-\frac{x}{b}\right) = -2b^2\exp\left(-\frac{x}{b}\right)\Big|_0^\infty = 2b^2\end{aligned}$$

Hence,

$$var[\gamma] = E[\gamma^2] - (E[\gamma])^2 = 2b^2 - 0 = 2b^2$$

3.

$$\begin{aligned}\Pr[|\gamma| > \lambda] = \Pr[\gamma > \lambda] + \Pr[\gamma < -\lambda] &= \int_\lambda^\infty \frac{1}{2b}\exp\left(-\frac{x}{b}\right)dx + \int_{-\infty}^{-\lambda} \frac{1}{2b}\exp\left(\frac{x}{b}\right)dx\\ &= 2\int_\lambda^\infty \frac{1}{2b}\exp\left(-\frac{x}{b}\right)dx = -\int_\lambda^\infty \exp\left(-\frac{x}{b}\right)d\left(-\frac{x}{b}\right) = -\exp\left(-\frac{x}{b}\right)\Big|_\lambda^\infty\\ &= \exp\left(-\frac{\lambda}{b}\right)\end{aligned}$$

■

Note that if b is small, the mass is highly concentrated around 0. Hence, it can only be used to privatize functions with small sensitivity. However, the highly concentrated mass also implies good utility. The following theorem shows that choosing $b := \frac{\Delta f}{\epsilon}$ is enough to preserve privacy and with high probability, the additive error incurred by the random noise is small.

Theorem 2.7 *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a deterministic function, $0 < \epsilon < 1$ be the privacy parameter and $0 < \delta < 1$ be the failure probability. Let $\gamma_1, \gamma_2, \ldots, \gamma_d$ be random variables independently sampled from $\mathsf{Lap}(\frac{\Delta f}{\epsilon})$. Then, the randomized function $\widehat{f}$ such that $\widehat{f}_i(X) := f_i(x) + \gamma_i$ for all $i \in [d]$*

1. *preserves ϵ-differential privacy,*
2. *is $(\frac{\Delta f}{\epsilon} \ln \frac{d}{\delta}, \delta)$-useful with respect to f.*

Proof: Let $X^{(1)} \in \mathcal{D}$ and $X^{(2)} \in \mathcal{D}$ be two neighboring database. Let $z \in \mathbb{R}^d$ be a vector. We abuse the notation a little bit and use $\Pr[\widehat{f}(X^{(1)}) = z]$ and $\Pr[\widehat{f}(X^{(2)}) = z]$ to denote the density instead of the probability. Hence, we have

$$\begin{aligned}
\frac{\Pr[\widehat{f}(X^{(1)}) = z]}{\Pr[\widehat{f}(X^{(2)}) = z]} &= \frac{\Pr[\wedge_{i=1}^{d} f_i(X^{(1)}) + \gamma_i = z_i]}{\Pr[\wedge_{i=1}^{d} f_i(X^{(2)}) + \gamma_i = z_i]} \\
&= \frac{\Pr[\wedge_{i=1}^{d} \gamma_i = z_i - f_i(X^{(1)})]}{\Pr[\wedge_{i=1}^{d} \gamma_i = z_i - f_i(X^{(2)})]} \\
&= \frac{\prod_{i=1}^{d} \Pr[\gamma_i = z_i - f_i(X^{(1)})]}{\prod_{i=1}^{d} \Pr[\gamma_i = z_i - f_i(X^{(2)})]} \\
&= \frac{\prod_{i=1}^{d} \frac{\epsilon}{2\Delta f} \exp(-\frac{|f_i(X^{(1)}) - z_i|}{\Delta f/\epsilon})}{\prod_{i=1}^{d} \frac{\epsilon}{2\Delta f} \exp(-\frac{|f_i(X^{(2)}) - z_i|}{\Delta f/\epsilon})} \\
&= \exp\left(\sum_{i=1}^{d} \left(\frac{|f(X^{(2)}) - z_i|}{\Delta f/\epsilon} - \frac{|f(X^{(1)}) - z_i|}{\Delta f/e} \right) \right) \\
&\le \exp\left(\sum_{i=1}^{d} \frac{|f(X^{(1)}) - f(X^{(2)})|}{\Delta f/\epsilon} \right) \\
&\le \exp\left(\frac{\Delta f}{\Delta f/\epsilon} \right) \\
&= \exp(\epsilon)
\end{aligned}$$

For any measurable subset $S \subseteq \mathbb{R}^d$, $\Pr[\widehat{f}(X^{(1)}) \in S] = \int_S \Pr[\widehat{f}(X^{(1)}) = z] dz \le \exp(\epsilon) \int_S \Pr[\widehat{f}(X^{(2)}) = z] dz = \Pr[\widehat{f}(X^{(2)}) \in S]$. Hence, the privacy guarantee is proved.

By Property 3 of Theorem 2.6, we know that for all $i \in [d]$, $\Pr[|\gamma_i| > \frac{\Delta f}{\epsilon} \ln \frac{d}{\delta}] = \exp(-(\frac{\Delta f}{\epsilon} \ln \frac{d}{\delta})/(\frac{\Delta f}{\epsilon})) = \frac{\delta}{d}$. Hence, by union bound on $i \in [d]$, we know that $\Pr[\vee_{i \in [d]} |\gamma_i| > \frac{\Delta f}{\epsilon} \ln \frac{d}{\delta}] \le \delta$.

Let $X \in \mathcal{D}$ be any database. Note that $|f_i(X) - \widehat{f_i}(X)| = |\gamma_i|$ for all $i \in [d]$. Hence, by the union bound, $\Pr[\exists i \in [d], |f_i(X) - \widehat{f_i}(X)| > \frac{\Delta f}{\epsilon} \ln \frac{d}{\delta}] \leq \delta$, which is equivalent to $\Pr[\wedge_{i \in [d]} |f_i(X) - \widehat{f_i}(X)| \leq \frac{\Delta f}{\epsilon} \ln \frac{d}{\delta}] \geq 1 - \delta$. Thus, the utility guarantee is proved. ■

2.5 Example

Consider the previous example where there are n users, each of which holds a private bit from $\{0, 1\}$, and we would like to calculate how many of them hold a one. In other words, we would like to calculate $\sum_{i=1}^{n} X_i$, where X_i is the private bit of the i-th user. There is an *untrusted aggregator* who can calculate the sum of n real numbers. To compute the sum, the users need to send information about their private bit to the aggregator, and the aggregator then does the calculation for them. However, since the aggregator might be malicious, the users do not want the aggregator to know their real private bits. The protect the users' privacy, the information they send to the aggregator should be ϵ-differentially private.

Note that the sensitivity of each user's private bit is one. Hence, a user can first add Laplace noise sampled from $\mathsf{Lap}(\frac{1}{\epsilon})$ to their private bit, and then send the noisy bit to the aggregator. The aggregator releases the sum of the noisy bits as an estimate of the real sum. Observe that the output is the real sum plus n independent random variables sampled from $\mathsf{Lap}(\frac{1}{\epsilon})$. It can be shown that the additive error of this estimate is $O(\frac{\sqrt{n}}{\epsilon})$ with high probability. This is left as an exercise.

2.6 Exercises

1. In this question, we derive a measure concentration result for independent random variables drawn from Laplace distribution. We show that with high probability, the sum of independent Laplace random variables are concentrated around its mean, 0.

   We use moment generating functions in a Chernoff-like argument. Let $\gamma_1, \gamma_2, \ldots, \gamma_n$ be n independent random variables, where γ_i is sampled from $\mathsf{Lap}(b_i)$.

   (a) Prove that for each γ_i, the moment generating function is $E[\exp(h\gamma_i)] = \frac{1}{1-h^2 b_i^2}$, where $|h| < \frac{1}{b_i}$.

   (b) Show that $E[\exp(h\gamma_i)] \leq \exp(2h^2 b_i^2)$, if $|h| < \frac{1}{\sqrt{2} b_i}$.
   (Hint: for $|x| < \frac{1}{2}$, we have $\frac{1}{1-x} \leq 1 + 2x \leq \exp(2x)$.)

   (c) Let $b_M := \max_{i \in [n]} b_i$. Also, let $\nu \geq \sqrt{\sum_{i=1}^{n} b_i^2}$ and $0 < \lambda < \frac{2\sqrt{2}\nu^2}{b_M}$. Prove that $\Pr[|Y| > \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{8\nu^2}\right)$

   (d) Suppose $0 < \delta < 1$ and $\nu > \max\left\{\sqrt{\sum_{i=1}^{n} b_i^2}, b_M \sqrt{\ln \frac{2}{\delta}}\right\}$. Prove that $\Pr[|Y| > \nu\sqrt{8 \ln \frac{2}{\delta}}] \leq \delta$.

   (e) Prove that $\Pr[|Y| > \sqrt{8} \cdot \sqrt{\sum_{i=1}^{n} b_i^2} \cdot \ln \frac{2}{\delta}] \leq \delta$.