Concentration inequalities

See Terry Tao's blog,
math 254 A notes 1
concentration of measure

Recall Markov, Chebyshev, Chernoff

Progressively more assumptions, tighter concentration

Today:
- Prove Chernoff
- more general result: subgaussian / subexponential

Example:

$Y_i = \min_x h_i(x)$ in distinct elements

$\Pr[Y_i > \frac{t}{n}] \leq e^{-t}$

Said Chebyshev $\Rightarrow$ after $O(\frac{1}{\varepsilon^2})$ trials,
mean $Y_i = E[Y_i] \pm \varepsilon/n$ w/ $\frac{3}{4}$ prob.

what about high probability? δ failure.

Chebyshev: $\frac{1}{\varepsilon^2 \delta}$

Chernoff? Y not bounded. –
just get median $O(\log \frac{1}{\delta})$ mean $O(1/\varepsilon^2)$ Y_i w/ $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$

today: mean$\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ works too.

Suppose X is mean zero.

$X = \sum_{i=1}^n X_i$, $X_i \in [-1, 1]$ iid

Chebyshev proof:

$$\Pr[|X| > t] = P[X^2 > t^2] \le \frac{E[X^2]}{t^2}.$$

$$E[X^2] = \underbrace{\sum_{i=1}^n E[X_i^2]}_{\le n} + \underbrace{\sum_{i\ne j} E[X_i X_j]}_{= E[X_i]E[X_j] = \text{zero}}$$

$$\Rightarrow \Pr[|X| > t] \le \frac{n}{t^2}.$$

$$\Pr[|X| > t\sqrt{n}] \le \frac{1}{t^2}$$

Why stop at 2nd moment?

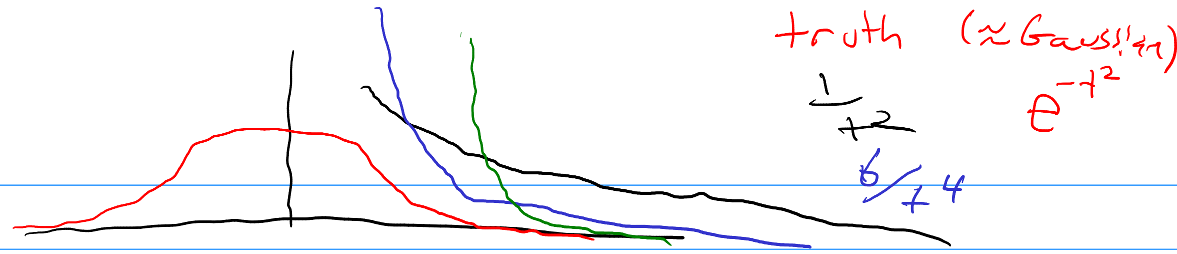
$$\Pr[|X| > t] = \Pr[X^4 > t^4] \le \frac{E[X^4]}{t^4}$$

$$E[X^4] = E\Big[\sum X_i^4 + \sum_{i\ne j} \binom{4}{2} X_i^2 X_j^2 + \cancel{\sum_{i \ne j} 4 X_i X_j^3}$$
$$+ \cancel{\sum_{i\ne j\ne k} \binom{4}{2} X_i^2 X_j X_k} + \cancel{\sum_{i\ne j\ne k\ne l} X_i X_j X_k X_l}\Big]$$

$\le 6n^2$.

$$\Pr[|X| > t\sqrt{n}] \le 6/t^4$$

Better tail, worse close in.



$\binom{k}{k/2}\cdot (k/2)! \approx \frac{2^k}{\sqrt{k}}\cdot \left(\frac{k}{2e}\right)^{k/2}\sqrt{2\pi n} \le k^{k/2}\cdot \text{const}$

k^{th} moment, k even:

Exercise: $E[x^k] \lesssim k^{k/2}\cdot n^{k/2}$

$\Rightarrow \Pr[|x| > t\sqrt{n}] \lesssim \left(\frac{\sqrt{k}}{t}\right)^k \quad \forall k$

For each t, optimize over k.

[take log, set $\frac{d}{dk} = 0$: $\cancel{\frac{1}{2}}\log k + \cancel{\frac{1}{2}} + \log\frac{e}{t^2} = 0 \Rightarrow k = \frac{t^2}{e}$]

$\Pr[|x| > t\sqrt{n}] \le e^{-t^2/e}$

Compare to our Chernoff:

$[-1,1] \to [0,1]$ by $y_i = \frac{x_i+1}{2}$.

Here: $\Pr[|y-\mu| > t\sqrt{n}] \lesssim e^{-4t^2/e}$

Claimed: $\Pr[|y-\mu| > \varepsilon\mu] \le 2e^{-\frac{\varepsilon^2}{2+\varepsilon}\mu}$

$t = \frac{\varepsilon\mu}{\sqrt{n}} \Rightarrow e^{-\frac{\varepsilon^2\mu^2}{n}\cdot 4/e} = e^{-\frac{\varepsilon^2\mu}{2}\cdot\frac{8}{e}\frac{\mu}{n}}$

$\Rightarrow$ what we claimed if $\mu \ge \frac{e}{8}n$.

More needed if $\mu \ll n$.

Idea: $\mu \ll n \Rightarrow x_i$ heavily biased $\Rightarrow E[x_i^2] \ll 1$.

$\Rightarrow E[x^k]$ smaller.

but helpful w/ limited independence

Moment calculations are kind of nasty.
Alternative: Moment Generating Function (MGF) $E[e^{\lambda X}]$

$$\Pr[X > t] \le \frac{E[e^{\lambda X}]}{e^{\lambda t}} \quad \forall \lambda > 0$$

($x < t$: use $\lambda < 0$)

Example: Gaussian, $p(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}}$

$$E[e^{\lambda X}] = \int_{-\infty}^{\infty} p(t) e^{\lambda t} dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t^2 - 2\sigma^2 \lambda t)} dt$$

$$= \int_{-\infty}^{\infty} p(t - \sigma^2\lambda) \cdot e^{\lambda^2\sigma^2/2} dt = e^{\lambda^2\sigma^2/2}$$

$$\Rightarrow \Pr[x > t] \le e^{\frac{\lambda^2\sigma^2}{2} - \lambda t} \qquad = e^{\frac{1}{2}(\lambda\sigma - \frac{t}{\sigma})^2} \cdot e^{\frac{-t^2}{2\sigma^2}}$$

optimize over λ

$$\Pr[x > t] \le e^{-\frac{t^2}{2\sigma^2}}$$

Definition

A mean zero random variable X is <u>subgaussian with parameter σ^2</u> if

$$E[e^{\lambda x}] \le e^{\lambda^2\sigma^2/2} \quad \forall \lambda.$$

In general, X is subgaussian w/ parameter σ^2 if $X - E[X]$ is.

Theorem For mean zero X, the following are equivalent up to changing σ by constant factors

(I) MGF: $E[e^{\lambda x}] \leq e^{\lambda^2 \sigma^2/2}$ $\forall \lambda$
(II) tail: $Pr[|x|>t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$ $\forall t>0$
(III) moments: $E[|x|^k] \leq \sigma^k \, k^{k/2}$ $\forall k$

As written, $I \Rightarrow II, III$ w/out changing constants

Lemma: Bounded variables are subgaussian.
If $X \in [a,b]$ then X is subgaussian$\left(\left(\frac{b-a}{2}\right)^2\right)$

$O\left(\left(\frac{b-a}{2}\right)^2\right)$ is trivial by Theorem, e.g. III.

Lemma: if X and Y independent, each subgaussian w/ σ_1^2 and σ_2^2, then $X+Y$ is subgaussian$(\sigma_1^2+\sigma_2^2)$

Pf $E[e^{\lambda(X+Y)}] = E[e^{\lambda x}]E[e^{\lambda y}] \leq e^{\frac{\lambda^2\sigma_1^2}{2}} e^{\frac{\lambda^2\sigma_2^2}{2}}$ ∎

Hence, if $X_1, \dots, X_n \in [0,1]$, we have:
X_i = subgaussian$\left(\left(\frac{1}{2}\right)^2\right)$
$X = \sum X_i$ = subgaussian$\left(\frac{n}{4}\right)$
$pr[|X-\mu|>t] \leq 2e^{-2t^2/n}$

[compare to moment bound: $e^{-\frac{4}{e}t^2/n}$]

Benefits: still works if X_i are heterogenous (some are larger) or Gaussian.
Still: not what was promised.

Two things we want:

(1) Concentrate mean(Y_i) where $P[Y_i > t] < e^{-t}$

(2) do better on $\sum X_i$ when $X_i \in [0,1]$ have $\mu_i << 1$

Problem: Y isn't subgaussian: e^{-t} not e^{-t^2}.

Suppose $p(z) = e^{-z}$ $\forall\ z \geq 0$.

$E[z] = 1$

MGF is

$$E[e^{\lambda(z - \frac{1}{e})}] = \int_0^\infty e^{-z} \cdot e^{\lambda z - \lambda}\, dz$$

$$= \frac{e^{-\lambda}}{1-\lambda} \qquad \text{for } \lambda < 1.$$

$$= \frac{1 - \lambda + \frac{\lambda^2}{2} - \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} \cdots}{1-\lambda} = 1 + \frac{\lambda^2}{2} + \lambda^3\left(\frac{1}{2} - \frac{1}{3!}\right) + \lambda^4\left(\frac{1}{4!} - \frac{1}{3!}\right)e \cdots$$

NB changed from last instantiation

Definition a mean zero X is subexponential w/ parameter σ if:

(I) $E[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$ $\forall\ |\lambda| < \frac{1}{\sigma}$

(II) $Pr[|X| > t] \leq 2e^{-\frac{t}{2\sigma}}$

(III) $E[|X|^k] \leq K^k \sigma^k$

which are equivalent up to constant factors in σ.

Unlike subgaussian, not closed under addition.

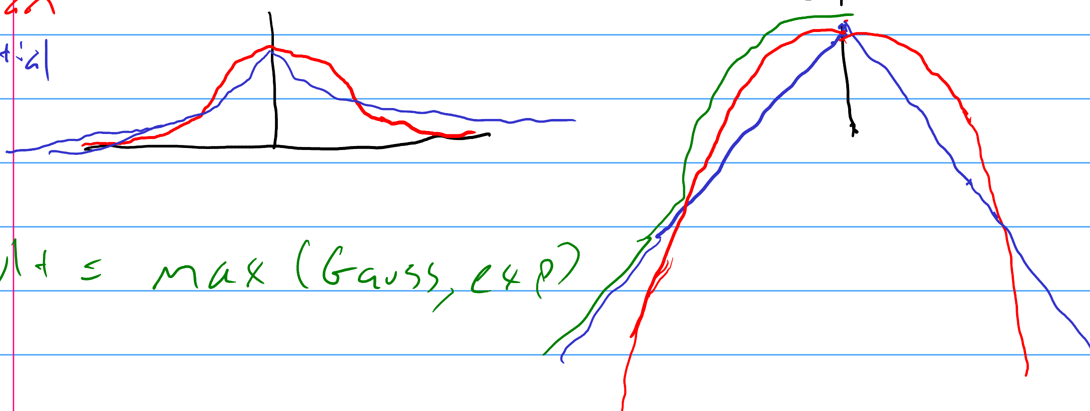
Definition A mean zero X is **Subgamma** w/ parameters (σ^2, c) if

$$\text{(I)} \quad E[e^{\lambda X}] \le e^{\frac{\lambda^2 \sigma^2}{2}} \quad \forall\ |\lambda| \le 1/c$$

This implies, and is equivalent up to constants to,

$$\Pr[x > t] \le \begin{cases} e^{-\frac{t^2}{2\sigma^2}} & \text{if } |t| \le \sigma^2/c \quad \xleftarrow{\lambda = \frac{t}{\sigma^2}} \\ e^{-\frac{t}{2c}} & \text{otherwise} \quad \xleftarrow{\lambda = \frac{1}{c}} \end{cases}$$

p $\qquad$ $\log p = e^{\max(-(\frac{t}{\sigma})^2, \frac{-t}{c})/2}$



result = max(Gauss, exp)

Gaussian interior, exponential tails

With pr. $1-\delta$, $|x| \le \max(\sigma\sqrt{2\log\frac{2}{\delta}},\ c \cdot 2\log\frac{2}{\delta})$

Intuition: Let Y_i = # coin flips till first heads

$Y = \sum_{i=1}^{n} Y_i$. $\qquad E[Y_i] = 2 \Rightarrow E[Y] = 2n$

$\qquad Var(Y) = n$

Central limit: converges to Gaussian,

If held everywhere, $\Pr[Y > 2n + t] \le e^{-\frac{t^2}{2n}}$

$\Pr[Y > 3n] \le e^{-n/2}$

$\Pr[Y > n^2] \le e^{-\Omega(n^3)}$

but $\Pr[Y > t] \ge \Pr[Y_1 > t] = 2^{-t}$.

$$\text{subexp}(\sigma) = \text{subgamma}(\sigma^2, \sigma)$$

$$X \in \text{subgamma}(s_1^2, c_1), \quad Y \in \text{subgamma}(s_2^2, c_2)$$
$$\Rightarrow X + Y \in \text{subgamma}(s_1^2 + s_2^2, \max(c_1, c_2))$$

Coin flip: $\text{subexp}(\theta(1))$

$$\Rightarrow Y \in \text{subgamma}(\theta(n), \theta(1))$$
$$\Rightarrow \Pr\left[Y > 2n + C\max\left(\sqrt{n \log \tfrac{1}{\delta}}, \log \tfrac{1}{\delta}\right)\right] \leq \delta$$

So indeed, Gaussianity stops at $2n + \theta(n)$.

Back to distinct elements $Y_i \in \text{subexp}(\sigma = \frac{1}{n})$

$$\Rightarrow \sum_{i=1}^{m} Y_i \in \text{subgamma}\left(\sqrt{\tfrac{m}{n}}^2, \tfrac{1}{n}\right)$$

$$\Rightarrow \frac{1}{m} \sum Y_i \in \text{subgamma}\left(\frac{1}{\sqrt{mn}}^2, \frac{1}{mn}\right)$$

$$\Pr\left[\left|\frac{1}{m} \sum Y_i - \text{mean}\right| > \frac{\varepsilon}{n}\right]$$
$$\leq 2 e^{-\frac{1}{2} \min\left(\frac{\varepsilon^2/n}{1/(nm)}, \frac{\varepsilon/n}{1/nm}\right)}$$
$$= 2 e^{-\frac{1}{2} \min(m\varepsilon^2, m\varepsilon)} = 2 e^{-\frac{1}{2} m \varepsilon^2}$$

$m = \frac{2}{\varepsilon^2} \log \frac{2}{\delta}$ suffices

How about $\sum X_i$ when $X_i \in [0,1]$, $E[X_i] = \mu$

Bernstein-type inequality.

Theorem: Let $X \in [0,1]$ have variance σ^2. Then $X \in \text{subgamma}(2\sigma^2, 2)$.

Pf/ Let $Y = X - E[X]$. Then

$$E[e^{\lambda Y}] = \sum_{k=0}^{\infty} \frac{\lambda^k E[Y^k]}{k!}$$

$$= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k E[Y^k]}{k!}$$

$$\boxed{E[Y^k] \leq E[Y^2 \cdot 1^{k-2}] = \sigma^2}$$

$$\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \sigma^2}{k!}$$

$$\leq 1 + \frac{\lambda^2 \sigma^2}{2} \cdot \left(\sum_{k=0}^{\infty} \lambda^k \right)$$

$$= 1 + \frac{\lambda^2 \sigma^2}{2(1-\lambda)}$$

For $\lambda < \frac{1}{2}$, then, $E[e^{\lambda Y}] \leq 1 + \lambda^2 \sigma^2$

$$\leq e^{\lambda^2 \sigma^2}.$$

$\Rightarrow$ subgamma$(2\sigma^2, 2)$.

Now, suppose $X = \sum x_i$, $x_i \in [0,1]$.
$E[x_i] = \mu_i$

$Var(x_i) = E[x_i^2] - \mu_i^2 \leq \mu_i - \mu_i^2 \leq \mu_i$

Each x_i is subgamma$(2\mu_i, 2)$

$\Rightarrow \sum x_i$ is subgamma$(2\mu, 2)$

$\Rightarrow \Pr[|\sum x_i - \mu| > t] \leq 2e^{-\frac{1}{2}\min\left(\frac{t^2}{2\mu}, \frac{t}{2}\right)}$

$\Pr[|\sum x_i - \mu| > \varepsilon\mu] \leq 2e^{-\frac{1}{4}\min(\varepsilon^2\mu, \varepsilon\mu)}$

Same as claimed in chernoff, up to constants.

$$\left[\min(\varepsilon^2, \varepsilon) = \theta\left(\frac{\varepsilon^2}{2+\varepsilon}\right) \right]$$

Note: stronger forms exist.

$\Pr[X > \frac{n}{2}]$ when $\mu_i = p$, binary.

Here: $e^{-\Omega(n)}$. Instead,

$\binom{n}{n/2} \cdot p^{n/2} \leq 2^n \cdot p^{n/2} = (4p)^n$

$= e^{-\Omega(n \log \frac{1}{p})}$

"Bennett's inequality" generalizes this