6.7720/18.619/15.070 Lecture 8
Sub-Exponential Random Variables and the Poisson Limit

Kuikui Liu

February 26, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials created by Guy Bresler and David Gamarnik, who taught previous iterations of this course. We are grateful for the discussions we had with them. We also consulted materials by Phillipe Rigollet, as well as "High-Dimensional Probability" by Roman Vershynin. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Weaknesses of Chernoff–Hoeffding

In the previous lecture, we looked at sub-Gaussian random variables and established various tail bounds for them. However, many important and natural random variables are emphatically *not* sub-Gaussian. Simple examples include the *Laplace (or symmetric exponential) distribution* $\mathsf{Lap}(\lambda)$ for $\lambda > 0$, which has density $x \mapsto \frac{\lambda}{2} e^{-\lambda \cdot |x|}$ for $x \in \mathbb{R}$, and the *Poisson distribution* $\mathsf{Poi}(\lambda)$ with mean $\lambda > 0$, which has probability mass function $k \mapsto \frac{\lambda^k e^{-\lambda}}{k!}$ for $k \in \mathbb{N}$. Yet, we still would like general-purpose concentration inequalities which apply to such random variables.

To illustrate the weakness of Chernoff–Hoeffding for such random variables, consider a collection of n independent Bernoulli random variables $X_1, \dots, X_n \sim \mathsf{Ber}(p)$ where $p = d/n$ for a constant d independent of n. Their sum $S_n = X_1 + \cdots + X_n$ is distributed as a binomial distribution $\mathsf{Bin}(n, d/n)$, which as we will show later in this lecture, converges to $\mathsf{Poi}(d)$ in the limit as $n \to \infty$; the Central Limit Theorem doesn't apply because the distribution of each X_i depends on n. As we previously mentioned, S_n has the same law as the degree of a vertex in the sparse Erdös–Rényi random graph $\boldsymbol{G}(n+1, d/n)$.

Now if we apply Chernoff–Hoeffding to S_n, we would obtain

$$\Pr\left[S_n - d \geq t\right] \leq \exp\left(-\frac{2t^2}{n}\right).$$

This tells us that S_n is at most $d + O\left(\sqrt{n}\right)$ with constant probability, which is useless since d is constant independent of n. Chebyshev already tells us that $S_n \leq d + O\left(\sqrt{d}\right)$ with 99% probability, since $\mathrm{Var}\left(S_n\right) = d\left(1 - \frac{d}{n}\right)$. The correct behavior is captured by the tails of $\mathsf{Poi}(d)$. One reason Chernoff–Hoeffding fails is that it only uses independence and *boundedness* of the random variables $X_1, \dots, X_n$; it doesn't use the crucial fact that most of the X_i are typically 0.

Notably, the situation cannot be remedied by replacing boundedness of the X_i with their variance proxy/sub-Gaussian norm. For $\mathsf{Ber}(d/n)$, one can see this via a direct calculation, but this is a symptom of a more general phenomenon. Namely, for distributions with tails as heavier than Gaussians, the moment generating function can be rather ill-behaved. To illustrate an extreme example, observe that the moment generating function for the Laplace distribution $\mathsf{Lap}(\lambda)$ with density $x \mapsto \frac{\lambda}{2} e^{-\lambda \cdot |x|}$ is

$$\mathbb{E}_{X \sim \mathsf{Lap}(\lambda)}\left[\exp\left(s \cdot X\right)\right] = \frac{1}{1 - \left(\frac{s}{\lambda}\right)^2}, \qquad \forall\, |s| < \lambda,$$

which has poles at $s = \pm\lambda$. Hence, one cannot hope to bound such a function by a function of the form $\exp\left(K^2 s^2\right)$, which is finite for all $s \in \mathbb{R}$.

2 Sub-Exponential Random Variables

To capture random variables with heavier tails, we define the class of *sub-exponential random variables*. Like the sub-Gaussian case, there are multiple equivalent definitions.

Proposition 2.1 (Sub-Exponential Distributions; see e.g. [Ver18])**.** *We say a random variable X is* sub-exponential *if it satisfies any one of the following definitions, which are all equivalent up to rescaling the corresponding factors $K_1, \dots, K_5$ by universal numerical constants:*

1. *The tails of X are upper bounded by exponential tails, i.e. there exists $K_1 > 0$ such that*

$$\Pr\left[|X| \geq t\right] \leq 2\exp\left(-t/K_1\right), \qquad \forall t \geq 0.$$

2. *The (absolute) moments of X are uniformly bounded by the moments of an exponential distribution, i.e. there exists $K_2 > 0$ such that*

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq K_2 \cdot p.$$

3. *The moment generating function of $|X|$ is upper bounded by an exponential function, i.e. there exists $K_3 > 0$ such that*

$$\mathbb{E}\left(\exp\left(s\,|X|\right)\right) \leq \exp\left(K_3 s\right), \qquad \forall s \in \left[0, \frac{1}{K_3}\right].$$

4. *The moment generating function of $|X|$ is upper bounded at some point, i.e. there exists $K_4 > 0$ such that $\mathbb{E}\left[\exp\left(|X|/K_4\right)\right] \leq 2$.*

5. *The moment generating function of the centered random variable $X - \mathbb{E}[X]$ is upper bounded by the moment generating function of a Gaussian in an interval, i.e. there exists $K_5 > 0$ such that*

$$\mathbb{E}\left[\exp\left(s \cdot (X - \mathbb{E}[X])\right)\right] \leq \exp\left(K_5^2 s^2\right), \qquad \forall s \in \left[-\frac{1}{K_5}, \frac{1}{K_5}\right].$$

Remark 1. The fifth condition is eerily similar to the fifth condition for sub-Gaussianity. The only difference lies in the restriction on the range of s. Unlike the sub-exponential case, for sub-Gaussian random variables, we require the inequality to hold for all $s \in \mathbb{R}$. However, as we saw for the Laplace distribution, the moment generating function of a sub-exponential random variable can explode at two poles. Hence, we can only hope for a bound in a neighborhood of the origin.

As before, we omit the proof as it is conceptually straightforward and relies entirely on calculations. Given these equivalent definitions, we can also quantify "how sub-exponential" a random variable is.

Definition 1 (Sub-Exponential Norm)**.** *For a sub-exponential random variable X, we define its* sub-exponential norm *as the quantity*

$$\|X\|_{\psi_1} \overset{\text{def}}{=} \inf\left\{K > 0 : \mathbb{E}\left[\exp\left(|X|/K\right)\right] \leq 2\right\}.$$

Lemma 2.2. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover, $\|X\|_{\psi_2}^2 = \left\|X^2\right\|_{\psi_1}$.*

Proof. This is immediate from the definitions. □

3 Concentration for Sub-Exponential Processes

Now let us establish a concentration inequality for linear functionals of sub-exponential processes.

Theorem 3.1 (Bernstein's Inequality)**.** *There exists a universal numerical constant $C > 0$ such that for any collection $X_1, \dots, X_n$ of independent sub-exponential random variables and any $\boldsymbol{v} \in \mathbb{R}^n$, the random variable $Y = \sum_{i=1}^n \boldsymbol{v}_i X_i$ satisfies the tail bound*

$$\Pr\left[Y - \mathbb{E}[Y] \geq t\right] \leq \exp\left(-\frac{1}{2C} \cdot \min\left\{\frac{t^2}{2C \cdot \sum_{i=1}^n \boldsymbol{v}_i^2 \left\|X_i\right\|_{\psi_1}^2}, \frac{t}{\max_{i=1,\dots,n} \left|\boldsymbol{v}_i\right| \cdot \left\|X_i\right\|_{\psi_1}}\right\}\right).$$

Very roughly speaking, one way to parse this bound is that there are two possible ways Y could deviate significantly from its expectation. The first possibility is that the "local deviations" $X_i - \mathbb{E}[X_i]$ all have the same sign, which then aggregate into a large deviation for Y. However, independence of the random variables makes this unlikely, leading to the first Gaussian-like term. A second possibility is that a single "local deviation" $X_i - \mathbb{E}[X_i]$ is large and contributes the bulk of the overall deviation $Y - \mathbb{E}[Y]$. This is "only exponentially unlikely" because we assumed the X_i are sub-exponential, not sub-Gaussian, and so this gives rise to the second exponential term in the minimum.

Proof of Theorem 3.1. As before, let us bound the moment generating function of $Y - \mathbb{E}[Y]$. By Proposition 2.1, there exists a universal numerical constant $c > 0$ such that for

$$K \stackrel{\mathsf{def}}{=} \max_{i=1,\dots,n} \left\{\left|\boldsymbol{v}_i\right| \cdot \left\|X_i\right\|_{\psi_1}\right\}$$
$$\widehat{\sigma}^2 \stackrel{\mathsf{def}}{=} \sum_{i=1}^n \boldsymbol{v}_i^2 \left\|X_i\right\|_{\psi_1}^2,$$

and any $s \in \left[-\frac{1}{C \cdot K}, \frac{1}{C \cdot K}\right]$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(s \cdot (Y - \mathbb{E}[Y])\right)\right] &= \prod_{i=1}^n \exp\left(s \cdot \boldsymbol{v}_i \cdot (X_i - \mathbb{E}[X_i])\right) && \text{(Independence)} \\
&\leq \prod_{i=1}^n \exp\left(C^2 \boldsymbol{v}_i^2 \left\|X_i\right\|_{\psi_1}^2 s^2\right) && \text{(Using Proposition 2.1)} \\
&= \exp\left(C^2 \widehat{\sigma}^2 s^2\right).
\end{aligned}$$

It follows by Markov's Inequality that

$$\begin{aligned}
\Pr\left[Y - \mathbb{E}[Y] \geq t\right] &= \inf_{s \in \left[-\frac{1}{C \cdot K}, \frac{1}{C \cdot K}\right]} \Pr\left[\exp\left(s \cdot (Y - \mathbb{E}[Y])\right) \geq e^{s \cdot t}\right] \\
&\leq \inf_{s \in \left[-\frac{1}{C \cdot K}, \frac{1}{C \cdot K}\right]} \exp\left(-st + C^2 \widehat{\sigma}^2 s^2\right).
\end{aligned}$$

The infimum is attained at $s = \min\left\{\frac{t}{2c^2 \widehat{\sigma}^2}, \frac{1}{c \cdot K}\right\}$. If t is such that $s = \frac{t}{2c^2 \widehat{\sigma}^2}$, then we obtain the first bound of $\exp\left(-\frac{t^2}{4C^2 \widehat{\sigma}^2}\right)$ as usual. On the other hand, if t is such that $s = \frac{1}{C \cdot K}$, then $\frac{t}{2C^2 \widehat{\sigma}^2} \geq \frac{1}{C \cdot K}$, or equivalently, $\frac{\widehat{\sigma}^2}{K^2} \leq \frac{t}{2C \cdot K}$. Plugging in $\frac{1}{C \cdot K}$ into $-st + C^2 \widehat{\sigma}^2 s^2$ gives $-\frac{t}{C \cdot K} + \frac{\widehat{\sigma}^2}{K^2} \leq -\frac{t}{2C \cdot K}$, which then yields the second bound of $\exp\left(-\frac{t}{2C \cdot K}\right)$. □

We highlight another version of Bernstein's Inequality for bounded random variables, whose proof we omit (note that the extra constant factors depending on $C > 0$ can be removed using more optimized arguments).

Theorem 3.2. *Let $X_1, \dots, X_n$ be independent mean-zero random variables which are bounded in the interval $[-K, K]$ almost surely. If $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \mathrm{Var}\left(S_n\right) = \sum_{i=1}^n \mathrm{Var}\left(X_i\right)$, then*

$$\Pr\left[S_n \geq t\right] \leq \exp\left(-\frac{t^2/2}{\sigma^2 + (Kt/3)}\right).$$

Let us return to the case where $X_1, \dots, X_n$ are i.i.d. $\mathsf{Ber}(d/n)$ random variables. We may apply Theorem 3.2 to obtain that for $S_n \sim \mathsf{Bin}(n, d/n)$,

$$\Pr\left[\left|S_n - d\right| \geq t\right] \leq 2\exp\left(-\frac{t^2/2}{d + (t/3)}\right).$$

This is almost the correct behavior. Indeed, the above is meaningful even for $t = \Theta\left(\sqrt{d}\right)$. Moreover, if we plug in $t = \alpha d$ for constant α, then the above yields an upper bound of $\exp\left(-C\alpha d\right)$ on the probability, while the correct Poisson tail yields $\exp\left(-\alpha d \log \alpha\right)$.

3.1 Thin Shell Concentration

Let us now use Bernstein's Inequality to study the norm of a sub-Gaussian random vector. In the previous lecture, we saw that if the entries are independent, mean-zero and $O(1)$-sub-Gaussian, then the norm is at most $O\left(\sqrt{n}\right)$ with overwhelming probability. Here, we show extremely strong concentration around $\sqrt{n}$, provided the entries are normalized to have unit variance.

Proposition 3.3. *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero unit-variance sub-Gaussian entries. Then there exists a universal numerical constant $C > 0$ such that for $\widehat{\sigma}^2 = \max_{i=1,\dots,n} \|X_i\|_{\psi_2}^2$,*

$$\left\| \|X\|_2 - \sqrt{n} \right\|_{\psi_2} \leq C\widehat{\sigma}^2.$$

Before we prove this, think about what this statement implies. It means that if the entries of X are $O(1)$-sub-Gaussian (e.g. you sample $X \sim \mathcal{N}(0, I_n)$), then the fluctuations of $\|X\|_2$ around $\sqrt{n}$ is of size $O(1)$, with no dependence on n whatsoever! In particular, 99.99% of the probability mass of a sub-Gaussian distribution is contained in a *thin shell* of constant thickness:

$$\left\{ \boldsymbol{x} \in \mathbb{R}^n : \sqrt{n} - c \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n} + c \right\}.$$

At first glance, this may seem like impossibly strong concentration. However, here's a simple back-of-the-envelope calculation which demystifies this. We know that $\|X\|_2^2$ has mean n, and a quick calculation leveraging independence and sub-Gaussianity shows its standard deviation is of order $\sqrt{n}$.[1] Hence, writing $x = y \pm \delta$ to mean $x \in [y - \delta, y + \delta]$, we have

$$\|X\|_2^2 \text{“}=\text{”} n \pm O\left(\sqrt{n}\right) \implies \|X\|_2 \text{“}=\text{”} \sqrt{n \pm O\left(\sqrt{n}\right)} \text{“}=\text{”} \sqrt{n} \pm O(1).$$

Remark 2. The famous *Thin Shell Conjecture* is asymptotic convex geometry asserts that for any *isotropic* (i.e. zero-mean, identity covariance) and *log-concave* probability measure μ on $\mathbb{R}^n$, the tails of $\left| \|X\|_2 - \sqrt{n} \right|$ are subexponential for $X \sim \mu$. In recent years, there has been a surge of progress on this and a whole host of interconnected conjectures. We refer interested readers to [Eld13; LV17; LV18; Che21; KL22; Gua24; KL24].

Proof of Proposition 3.3. The idea is that by assumption, for each $i = 1, \dots, n$, the random variable $X_i^2 - 1$ is sub-exponential, and so we can apply Bernstein's Inequality to get an exponential tail for the deviation of $\|X\|_2^2$ from n. Taking a square root then gives a sub-Gaussian tail. Indeed, if Y is a sub-exponential random variable, then

$$\Pr\left[\sqrt{|Y|} \geq t\right] = \Pr\left[|Y| \geq t^2\right] \leq 2\exp\left(-t^2 / \|Y\|_{\psi_1}\right),$$

so $\sqrt{|Y|}$ has a sub-Gaussian tail.

To formalize everything, we leverage an intermediate lemma.

Claim 3.4. *There exists a universal numerical constant $C > 0$ such that $\left\|X_i^2 - 1\right\|_{\psi_1} \leq C \|X_i\|_{\psi_2}^2$.*

Proof. Proposition 2.1 and the definition of $\|\cdot\|_{\psi_1}$ imply that $\left\|X_i^2 - 1\right\|_{\psi_1} \leq C \left\|X_i^2\right\|_{\psi_1}$ for some $C > 0$. Applying Lemma 2.2 completes the proof. □

Returning to the Euclidean norm, we combine Claim 3.4 with Bernstein's Inequality (see Theorem 3.1) to obtain

$$\begin{aligned}
\Pr\left[\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq t\right] &\leq 2\exp\left(-\frac{L \cdot n}{\widehat{\sigma}^2} \cdot \min\left\{\frac{t^2}{\widehat{\sigma}^2}, t\right\}\right) \\
&\leq 2\exp\left(-\frac{L' \cdot n}{\widehat{\sigma}^4} \cdot \min\left\{t^2, t\right\}\right)
\end{aligned}$$

for some universal numerical constants $L, L' > 0$; note that in the second step, we used the fact that $\mathrm{Var}(X_i) = 1$ for all i implies $\widehat{\sigma}^2$ is lower bounded by some universal numerical constant (see the previous lecture).

[1] Observe that $\mathrm{Var}\left(\|X\|_2^2\right) = \sum_{i=1}^n \mathrm{Var}\left(X_i^2\right) \leq O(n)$.

Let's now translate this into a bound on the tail of $\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right|$. Observe that if $x \geq 0$ satisfies $|x-1| \geq s$, then $\left|x^2 - 1\right| \geq \max\{s, s^2\}$; this can be verified by separately checking the case $x \leq \max\{0, 1-s\}$ and the case $x \geq 1+s$. It follows that

$$\begin{aligned}\Pr\left[\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq s\right] &\leq \Pr\left[\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq \max\{s, s^2\}\right] \\ &\leq 2\exp\left(-\frac{L' \cdot n}{\widehat{\sigma}^4} \cdot \underbrace{\min\left\{\max\{s^2, s^4\}, \max\{s, s^2\}\right\}}_{=s^2 \text{ (check } s>1 \text{ and } s<1 \text{ separately)}}\right) \\ &= 2\exp\left(-\frac{L' \cdot n}{\widehat{\sigma}^4} \cdot s^2\right).\end{aligned}$$

Substituting $t = s \cdot \sqrt{n}$ back in, we obtain

$$\Pr\left[\left|\|X\|_2 - \sqrt{n}\right| \geq t\right] \leq 2\exp\left(-O\left(\frac{t^2}{\widehat{\sigma}^4}\right)\right), \qquad \forall t \geq 0.$$

This implies that $\left\|\|X\|_2 - \sqrt{n}\right\|_{\psi_2} \lesssim \widehat{\sigma}^2$ as desired. □

4 The Poisson Limit of Sparse Binomials

To conclude this lecture, we prove that $\mathsf{Bin}(n, d/n)$ is approximately $\mathsf{Poi}(d)$ in the large n limit for constant d. To formalize this, let us recall that for two probability measures μ, ν on a common state space Ω, their total variation distance is

$$\|\mu - \nu\|_{\mathsf{TV}} \stackrel{\mathsf{def}}{=} \frac{1}{2}\sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|.$$

We bound the total variation distance between $\mathsf{Bin}(n, d/n)$ and $\mathsf{Poi}(d)$ by $O(1/n)$.

Theorem 4.1 (Binomial–Poisson Approximation)**.** *For any $d \in \mathbb{R}_{\geq 0}$ and $n \in \mathbb{N}$, we have*

$$\|\mathsf{Bin}(n, d/n) - \mathsf{Poi}(d)\|_{\mathsf{TV}} \leq \frac{d^2}{n}.$$

4.1 The Coupling Interpretation of Total Variation

We previously saw that total variation distance can be interpreted through the lens of test functions which "distinguish" μ from ν; more specifically, $\|\mu - \nu\|_{\mathsf{TV}} = \sup_{f:\Omega\to[0,1]} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]|$. Towards proving Theorem 4.1, we introduce a "dual" interpretation of the total variation distance based on *coupling*.

Definition 2 (Coupling)**.** *Let μ, ν be probability measures on Ω, Σ, respectively. A* coupling *of μ, ν is a probability measure ξ on $\Omega \times \Sigma$ such that*

$$\mu(x) = \sum_{y \in \Sigma} \xi(x, y), \qquad \forall x \in \Omega$$
$$\nu(y) = \sum_{x \in \Omega} \xi(x, y), \qquad \forall y \in \Sigma.$$

In other words, the marginals of ξ on each coordinate are precisely μ, ν, respectively.

One way to think about a coupling is through matrices. If we view μ, ν as vectors in $\mathbb{R}^\Omega$, then ξ is a matrix in $\mathbb{R}^{\Omega\times\Omega}$ such that the rows of ξ sum to μ, and the columns of ξ sum to ν.

Probabilistically, one should think of a coupling of μ, ν as a method for sampling a (possibly correlated, or "coupled") pair of random variables (X, Y) such that $\mathsf{Law}(X) = \mu$ (ignoring Y), and $\mathsf{Law}(Y) = \nu$ (ignoring X). Couplings always exist, since we always have the *product measure* $(\mu \otimes \nu)(x, y) \stackrel{\mathsf{def}}{=} \mu(x) \cdot \nu(y)$, where we simply sample $X \sim \mu, Y \sim \nu$ independently. If $\mu = \nu$, then we also have the *identity coupling*, where $\xi(x, x) = \mu(x) = \nu(x)$ for all x, and $\xi(x, y) = 0$ for all $x \neq y$. Algorithmically, we just sample $X \sim \mu$ and output two copies (X, X). Let us see how to optimally couple two coins with different biases.

Example 1. Fix $p, q \in [0,1]$, and assume $p \le q$ without loss of generality. We may couple $X \sim \mathsf{Ber}(p)$ and $Y \sim \mathsf{Ber}(q)$ as follows: First, sample $U \sim \mathsf{Unif}[0,1]$. Then, output

$$(X,Y) = \begin{cases} (1,1), & \text{if } 0 \le U \le p \\ (0,1), & \text{if } p < U \le q \\ (0,0), & \text{if } q < U \le 1 \end{cases}.$$

Clearly, $\Pr[X=1] = p$ and $\Pr[Y=1] = q$. When $p = q$, then $\Pr[X = Y] = 1$ and the two coins are maximally correlated. On the other hand, if $p \neq q$, then $\Pr[X \neq Y] = |p - q|$.

Example 1 suggests that couplings can be used to quantify the distance between laws of random variables. In fact, it turns out they are intimately connected to total variation distance. We have already seen two equivalent definitions of the latter; the following lemma furnishes a third equivalent definition.

Lemma 4.2 (Coupling Lemma)**.** *Let μ, ν be two probability distributions on a common state space Ω. Then for any coupling ξ of μ, ν,*

$$\|\mu - \nu\|_{\mathsf{TV}} \le \Pr_{(X,Y)\sim\xi}[X \neq Y].$$

Moreover, there exists a coupling, colloquially referred to as the TV-optimal coupling*, which achieves equality.*

Remark 3. The Coupling Lemma allows one to interpret the total variation distance as a *transportation distance*, which are fundamental to the theory of *optimal transport*. Furthermore, equality of the characterizations based on test functions and couplings can be interpreted through the lens of *linear programming duality*.

For the moment, we focus only on the upper bound, and defer the proof of TV-optimality to Appendix A.

Proof of the Upper Bound. Because ξ is a coupling, $\xi(x,x) \le \min\{\mu(x), \nu(x)\}$ for all $x \in \Omega$. It follows that

$$\Pr_{(X,Y)\sim\xi}[X = Y] = \sum_{x\in\Omega} \xi(x,x) \le \sum_{x\in\Omega} \min\{\mu(x), \nu(x)\}.$$

On the other hand,

$$\|\mu - \nu\|_{\mathsf{TV}} = \sum_{x:\mu(x)\ge\nu(x)} (\mu(x) - \nu(x)) = \sum_{x\in\Omega} (\mu(x) - \min\{\mu(x),\nu(x)\}) = 1 - \sum_{x\in\Omega} \min\{\mu(x),\nu(x)\}.$$

Combining the preceding two displays concludes the proof. □

4.2 Proof of Theorem 4.1

To bound the total variation distance, recall that $X \sim \mathsf{Bin}(n, d/n)$ can be decomposed as a sum of n independent random variables $X_1, \dots, X_n \sim \mathsf{Ber}(d/n)$. It turns out any Poisson random variable can be decomposed in a similar way.

Lemma 4.3. *If $Y_1 \sim \mathsf{Poi}(\lambda_1), Y_2 \sim \mathsf{Poi}(\lambda_2)$ are independent, then $Y_1 + Y_2 \sim \mathsf{Poi}(\lambda_1 + \lambda_2)$.*

In particular, we can decompose $Y \sim \mathsf{Poi}(d)$ as $Y = \sum_{i=1}^n Y_i$ for independent $Y_1, \dots, Y_n \sim \mathsf{Poi}(d/n)$. The proof of Lemma 4.3 is a straightforward exercise in calculus.

Now, observe that any coupling of $X_i \sim \mathsf{Ber}(d/n)$ with $Y_i \sim \mathsf{Poi}(d)$ for each $i = 1, \dots, n$ naturally induces a coupling of $X \sim \mathsf{Bin}(n, d/n), Y \sim \mathsf{Poi}(d)$. In particular, if we use the TV-optimal coupling between X_i, Y_i, then

$$\begin{aligned}
\|\mathsf{Bin}(n,d/n) - \mathsf{Poi}(d)\|_{\mathsf{TV}} &\le \Pr[X \neq Y] && \text{(Coupling Lemma; see Lemma 4.2)} \\
&\le \Pr[\exists i \text{ s.t. } X_i \neq Y_i] && \text{(If } X_i = Y_i \text{ for all } i\text{, then } X = Y\text{)} \\
&\le \sum_{i=1}^n \Pr[X_i \neq Y_i] && \text{(Union Bound)} \\
&= n \cdot \|\mathsf{Ber}(d/n) - \mathsf{Poi}(d/n)\|_{\mathsf{TV}} && \text{(TV-optimality of the coupling)}
\end{aligned}$$

Hence, it suffices to upper bound $\|\mathsf{Ber}(d/n) - \mathsf{Poi}(d/n)\|_{\mathsf{TV}}$. A short calculation employing the ℓ_1-viewpoint of total variation reveals that

$$\|\mathsf{Ber}(d/n) - \mathsf{Poi}(d/n)\|_{\mathsf{TV}} = \frac{d}{n}\left(1 - e^{-d/n}\right) \leq \frac{d^2}{n^2},$$

where in the final step, we used $1 - x \leq e^{-x}$ for $x \in \mathbb{R}$. A coupling of $\mathsf{Ber}(d/n), \mathsf{Poi}(d/n)$ achieving the above identity is given by

$$\Pr[X = x, Y = y] = \begin{cases} 1 - \frac{d}{n}, & \text{if } x = y = 0 \\ e^{-d/n} - \left(1 - \frac{d}{n}\right), & \text{if } x = 1, y = 0 \\ \frac{(d/n)^y e^{-d/n}}{y!}, & \text{if } x = 1, y \geq 1 \end{cases}.$$

Remark 4. This proof (and the statement of Theorem 4.1) can be easily generalized to bound the total variation distance between $\mathsf{Poi}\left(\sum_{i=1}^n p_i\right)$ and $\mathsf{Law}\left(\sum_{i=1}^n X_i\right)$, where $X_i \sim \mathsf{Ber}(p_i)$ are independently drawn. The corresponding bound becomes $\sum_{i=1}^n p_i^2$, a result known as *Le Cam's Inequality*.

References

[Che21] Yuansi Chen. "An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture". In: *Geometric and Functional Analysis* 31 (2021), pp. 34–61 (cit. on p. 4).

[Eld13] Ronen Eldan. "Thin Shell Implies Spectral Gap Up to Polylog via a Stochastic Localization Scheme". In: *Geometric and Functional Analysis* 23 (2013), pp. 532–569 (cit. on p. 4).

[Gua24] Qing-Yang Guan. "A note on Bourgain's slicing problem". In: *arXiv preprint arXiv:2412.09075* (2024) (cit. on p. 4).

[KL22] Bo'az Klartag and Joseph Lehec. "Bourgain's slicing problem and KLS isoperimetry up to polylog". In: *arXiv preprint arXiv:2203.15551* (2022) (cit. on p. 4).

[KL24] Boaz Klartag and Joseph Lehec. "Affirmative Resolution of Bourgain's Slicing Problem using Guan's Bound". In: *arXiv preprint arXiv:2412.15044* (2024) (cit. on p. 4).

[LV17] Yin Tat Lee and Santosh Srinivas Vempala. "Eldan's Stochastic Localization and the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. 2017, pp. 998–1007. DOI: 10.1109/FOCS.2017.96 (cit. on p. 4).

[LV18] Yin Tat Lee and Santosh S. Vempala. "Stochastic Localization + Stieltjes Barrier = Tight Bound for Log-Sobolev". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2018. Los Angeles, CA, USA: Association for Computing Machinery, 2018, pp. 1122–1129. ISBN: 9781450355599. DOI: 10.1145/3188745.3188866 (cit. on p. 4).

[Ver18] Roman Vershynin. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018 (cit. on p. 2).

A Existence of a TV-Optimal Coupling

To construct a TV-optimal coupling, note that any such coupling must saturate the only inequality we used in the proof of the upper bound in Lemma 4.2, i.e. we must ensure that $\xi(x,x) = \min\{\mu(x), \nu(x)\}$ for all $x \in \Omega$. At this point, I could just hand you a clean formula and save all of us some pain (see Remark 5). Here is how one could try to reason about it step by step. Let $A = \{x : \mu(x) > \nu(x)\}$, $B = \{x : \nu(x) > \mu(x)\}$ and $C = \{x : \mu(x) = \nu(x)\}$. Then, as a $\Omega \times \Omega$ matrix with rows summing to μ and columns summing to ν, we need ξ to have the following block structure:

1. Along the diagonal blocks $A \times A$, $B \times B$ and $C \times C$, we have diagonal entries $\xi(x,x) = \min\{\mu(x), \nu(x)\}$. The off-diagonal entries for these blocks must all be zero since one of the marginal distributions must be saturated.

2. The blocks $A \times C$, $B \times C$, $C \times A$ and $C \times B$ must all be zero both marginals μ, ν have already been saturated by the $C \times C$ block. Similarly, the $B \times A$ block must be zero.

$$\xi = \begin{array}{c} \begin{matrix} A & B & C \end{matrix} \\ \begin{pmatrix} \text{diag} & ??? & 0 \\ 0 & \text{diag} & 0 \\ 0 & 0 & \text{diag} \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \end{matrix}$$

Thus, the only freedom we have in choosing our optimal coupling is designing the $A \times B$ submatrix of ξ such that

$$\mu(x) - \nu(x) = \sum_{y \in B} \xi(x, y), \qquad \forall x \in A$$
$$\nu(y) - \mu(y) = \sum_{x \in A} \xi(x, y), \qquad \forall y \in B.$$

Since $\mu(A) - \nu(A) = \nu(B) - \mu(B)$, such a submatrix is always possible. For instance, one can sort A in increasing order of $\mu(x) - \nu(x)$, sort B analogously, and inductively build a triangular matrix.

Remark 5. Here is a formula for such an optimal coupling. Take $\xi(x, x) = \min\{\mu(x), \nu(x)\}$ for all $x \in \Omega$, and for $x \neq y$, set

$$\xi(x, y) = \frac{(\mu(x) - \xi(x, x)) \cdot (\nu(y) - \xi(y, y))}{1 - \sum_{z \in \Omega} \xi(z, z)}.$$