1 Introduction

A few times in this class we've used the fact that Lipschitz functions of independent random variables have tight (Gaussian) concentration.

Theorem 1 (McDiarmid's Inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables and f be a function of them such that changing X_i changes f by at most L_i. Then for all $\lambda \geq 0$,*

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq \lambda\right] \leq 2\exp\left(-\frac{\lambda^2}{2\|L\|_2^2}\right).$$

Proof. We use Chernoff's method, assuming without loss of generality that $\mathbb{E}[f(X)] = 0$:

$$\mathbb{P}[f(X) > \lambda] \leq e^{-\theta\lambda} m(\theta),$$

where $m(\theta) = \mathbb{E}[e^{\theta f(X)}]$. The main step is to prove the following bound:

$$m(\theta) \leq e^{\|L\|_2^2 \theta^2/2},$$

and then choosing $\theta = \frac{\lambda}{\|L\|_2^2}$ gives the desired result. The above bound is proved by considering the martingale $M_t = \mathbb{E}[f(X)|X_1, \dots, X_t]$, noting that the difference $M_t - M_{t-1}$ is bounded by L_t in absolute value, and then using Höfding's lemma which gives that

$$\mathbb{E}[e^{\theta(M_t - M_{t-1})}|X_1, \dots, X_{t-1}] \leq e^{\theta^2 L_t^2/2}.$$

So, splitting up the MGF into a telescoping product, we obtain the desired bound. ∎

The intuition here is that there is a fast sampling algorithm for X (namely, just sample each X_i one-by-one) such that the value, or expected value, of f doesn't change much at each step of the algorithm. It turns out there is a way to extend this to correlated random variables, as long as a reversible Markov chain with the given distribution as its stationary distribution mixes quickly.

2 Bounding the MGF

One alternate way to bound the MGF of $f(X)$ is to use a differential inequality. Suppose $\mathbb{E}[f(X)] = 0$ without loss of generality. Namely, if we can prove that $m(\theta) = \mathbb{E}[e^{\theta f(X)}]$ satisfies

$$m'(\theta) \leq C\theta \cdot m(\theta) \tag{1}$$

for $\theta \geq 0$, then we find that $m(\theta) \leq e^{C\theta^2/2}$ for $\theta \geq 0$, since $m(0) = 1$. This would then imply that

$$\mathbb{P}[f(X) \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2C}\right)$$

by Markov's Inequality.

How might we prove (1)? Since

$$m'(\theta) = \mathbb{E}[f(X)e^{\theta f(X)}],$$

we'd like to be able to somehow extract the $f(X)$ from the expression and pay a cost of $C\theta$. Let Y be obtained from X by taking one step in the Markov chain. By stationarity we have

$$\mathbb{E}[f(X)e^{\theta f(X)}] = \mathbb{E}[f(Y)e^{\theta f(Y)}].$$

Now if we had some function $F(x, y)$ so that

$$\mathbb{E}[F(X,Y)|X] = f(X),$$
$$\mathbb{E}[F(X,Y)|Y] = -f(Y),$$

then the above equation would yield

$$\begin{aligned} m'(\theta) &= \frac{1}{2}\left(\mathbb{E}[F(X,Y)e^{\theta f(X)}] - \mathbb{E}[F(X,Y)e^{\theta f(Y)}]\right) \\ &= \frac{1}{2}\mathbb{E}[F(X,Y)(e^{\theta f(X)} - e^{\theta f(Y)})]. \end{aligned}$$

Now we have a difference of the MGF at two different points, and we can use the following lemma:

Lemma 2. *For $a, b \in \mathbb{R}$,*

$$\frac{e^a - e^b}{a - b} \leq \frac{1}{2}(e^a + e^b).$$

Proof. We have the following integral formula:

$$\begin{aligned} \frac{e^a - e^b}{a - b} &= \int_0^1 e^{t(a-b)+b}\, dt \\ &= \int_0^1 e^{ta+(1-t)b}\, dt \\ &\leq \int_0^1 (te^a + (1-t)e^b)\, dt \\ &= \frac{1}{2}(e^a + e^b). \end{aligned}$$

The inequality is by the convexity of $u \mapsto e^u$. ∎

Note that the mean value theorem does not suffice here, the above lemma is crucial since we don't know which of $f(X)$ or $f(Y)$ is larger, and we don't want to consider things like $e^{\theta \max\{f(X), f(Y)\}}$, but rather things which we can reduce to $m(\theta)$ again. Anyway, the lemma gives

$$m'(\theta) \leq \frac{\theta}{4}\mathbb{E}[|F(X,Y)||f(X) - f(Y)|(e^{\theta f(X)} + e^{\theta f(Y)})], \tag{2}$$

and we have successfully extracted the θ and have something that can be massaged into another copy of $m(\theta)$, with some effort. Suppose that we have

$$|F(X,Y)||f(X) - f(Y)| \leq 2C$$

for all X, Y that are one step apart in the Markov chain. Then we get

$$m'(\theta) \leq C\theta \cdot m(\theta),$$

using the fact that X and Y have the same distribution. Actually, we can get away with the following seemingly weaker assumption:

$$\mathbb{E}\left[|F(X,Y)||f(X) - f(Y)| \,|\, X\right] \leq 2C \tag{3}$$

almost surely. This is because we can split up the expression (2) into two terms and use the fact that X and Y have the same distribution.

This finishes the proof of (1), apart from two things. First of all, what is the function F? Secondly, how do we prove (3)?

3 The function F

We want a function $F(x, y)$ which satisfies

$$\mathbb{E}[F(X, Y)|X] = f(X),$$
$$\mathbb{E}[F(X, Y)|Y] = -f(Y),$$

so let's start with $f(x) - f(y)$. But now

$$\mathbb{E}[f(X) - f(Y)|X] = f(X) - \mathbb{E}[f(Y)|X],$$
$$\mathbb{E}[f(X) - f(Y)|Y] = \mathbb{E}[f(X)|Y] - f(Y),$$

so we need to correct this. We can't just subtract off $f(X)$ and add $f(Y)$, but notice that if we instead run a new Markov chain (X^t) started at $X^0 = X$ and similarly for (Y^t), we have

$$\mathbb{E}[f(Y)|X] = \mathbb{E}[f(X^1)|X],$$
$$\mathbb{E}[f(X)|Y] = \mathbb{E}[f(Y^1)|Y],$$

using reversibility. So instead we should consider

$$f(x) - f(y) + \mathbb{E}[f(X^1)|X = x] - \mathbb{E}[f(Y^1)|Y = y].$$

But now, conditioned on X, we have the error term $-\mathbb{E}[f(Y^1)|X]$, which is the same as $-\mathbb{E}[f(X^2)|X]$, and conditioned on Y we have the error term $+\mathbb{E}[f(X^1)|Y]$, which is the same as $+\mathbb{E}[f(Y^2)|Y]$. Adding extra terms to cancel these, the pattern continues, and we are left with

$$F(x, y) \coloneqq \sum_{t=0}^{\infty} \mathbb{E}[f(X^t) - f(Y^t)|X = x, Y = y].$$

Note that this does not depend on the coupling between the chains (X^t) and (Y^t), and so we can choose an advantageous coupling to ensure that $F(X, Y)$ is small.

4 The assumption (3)

The construction of F is standard, but the task of verifying that assumption (3) will depend on the particular Markov chain we're working with, as well as the function f for which we'd like to show concentration at stationarity.

For concreteness, let's suppose the X_i are all $\{0, 1\}$-valued and that the function f is L-Lipschitz, where L is a vector, meaning that $|f(x) - f(y)| \leq L_i$ if x and y differ at the ith coordinate alone. Let us also suppose that we are in the Dobrushin contraction regime for whatever distribution we have on the hypercube, meaning that, with $d(x, y)$ denoting the Hamming distance between x and y,

$$\mathbb{E}[d(X^1, Y^1)] \leq \left(1 - \frac{c}{n}\right) \cdot d(X, Y)$$

for some $c > 0$, under the Glauber dynamics. This is the case with many Gibbs measures on the hypercube at high enough temperature. A strategy similar to the following will also work in any situation where the Markov chain mixes quickly, even outside of the Dobrushin regime, with perhaps a few extra steps related to a possible "burn-in" period.

Now, letting A_i denote the event that index i is chosen to get from X to Y, we have

$$\begin{aligned}\mathbb{E}\left[|f(X) - f(Y)||F(X, Y)||X\right] &\leq \frac{1}{n} \sum_{i=1}^{n} L_i \cdot \mathbb{E}\left[|F(X, Y)||A_i\right] \\ &\leq \frac{1}{n} \sum_{i=1}^{n} L_i \cdot \sum_{t=0}^{\infty} \mathbb{E}\left[\left|f(X^t) - f(Y^t)\right|\middle|A_i\right].\end{aligned}$$

Now, we have $|f(X^t) - f(Y^t)| \leq \|L\|_\infty \cdot d(X^t, Y^t)$ since each time a coordinate is changed to get from X^t to Y^t, the function can change by at most $\|L\|_\infty$. So, continuing the chain of inequalities, we have

$$\begin{aligned}
\mathbb{E}\left[|f(X) - f(Y)||F(X,Y)||X\right] &\leq \frac{1}{n}\sum_{i=1}^{n} L_i \cdot \sum_{t=0}^{\infty} \|L\|_\infty \cdot \mathbb{E}\left[d(X_j^t, Y_j^t)\middle|A_i\right] \\
&\leq \frac{\|L\|_\infty}{n}\sum_{i=1}^{n} L_i \cdot \sum_{t=0}^{\infty}\left(1 - \frac{c}{n}\right)^t \\
&\leq \frac{\|L\|_\infty}{n}\sum_{i=1}^{n} L_i \cdot \frac{n}{c} \\
&= \frac{1}{c}\|L\|_\infty\|L\|_1,
\end{aligned}$$

Therefore, by the general machinery, we obtain

$$\mathbb{P}[f(X) \geq \lambda] \leq \exp\left(-\frac{2c\lambda^2}{\|L\|_\infty\|L\|_1}\right).$$

For instance, if $M = 2\sum_{i=1}^n X_i - n$, i.e. the total magnetization in an Ising model, then we obtain

$$\mathbb{P}[|M| \geq \lambda] \leq 2\exp\left(-\frac{c\lambda^2}{2n}\right),$$

meaning we get Gaussian concentration at scale $\sqrt{n}$, the same as for a sum of i.i.d. Bernoulli random variables. Note that this recovers the optimal bound in the infinite temperature case, where all spins are independent, as there we have $c = 1$ and $\mathrm{Var}(M) = n$. So it is a safe bet that this is quite a good result in the general case, at least within the Dobrushin contraction regime.

5 A word on history

This method is often called "Stein's method for proving concentration," although it was actually developed by Chatterjee in his PhD thesis in 2005. Stein himself did use the idea of exchangeable pairs (one step of a reversible stationary Markov chain) to prove effective versions of the central limit theorem, for instance the following which may be useful to know:

Theorem 3. *Let X, Y be an exchangeable pair such that $\mathbb{E}[Y|X] = (1-a)X$ for $0 < a \leq 1$. Then*

$$d_{\mathrm{Was}}(X, \mathcal{N}(0,1)) \leq \frac{\sqrt{\mathrm{Var}(\mathbb{E}[(X-Y)^2|X])}}{\sqrt{2\pi}a} + \frac{\mathbb{E}[|X-Y|^3]}{3a}. \tag{4}$$

This was the first of many results of this type, collectively called *Stein's method*, wherein bounds on various natural distances between a distribution and the standard normal (or another target distribution) are obtained by bounding the so-called characterizing operator of the distribution on a natural class of functions. Specifically, for the standard normal, the characterizing operator is

$$\mathcal{A}f(x) = f'(x) - xf(x),$$

which satisfies $\mathbb{E}[\mathcal{A}f(Z)] = 0$ for all f when $Z \sim \mathcal{N}(0,1)$. The following theorem, from which the previous result can be derived, is a typical example of the foundational meta-theorems of Stein's method.

Theorem 4.

$$d_{\mathrm{Was}}(X, \mathcal{N}(0,1)) = \sup_{f \in \mathrm{Lip}^1} \left|\mathbb{E}[\mathcal{A}f(X)]\right|.$$

Notably, Stein did not use these ideas to prove concentration results, but was more focused on bounding natural distances to the normal distribution, which typically does not give full Gaussian concentration.