8
Gaussians, Martingales, and Discrepancy

8.1 Introduction

In this lecture, we explore the interplay between Gaussian random variables, dimension reduction, martingale concentration, and their application to the problem of discrepancy minimization.

The plan for today is:

1. Basics about Gaussian random variables (r.v.s).
2. Dimension reduction via the Johnson-Lindenstrauss Lemma.
3. A recap of martingale concentration, particularly for Gaussians.
4. Discrepancy minimization using Random Walks, Martingales, and Gaussians.

8.2 Facts about Gaussian Random Variables

We start by defining the Gaussian distribution and recalling some fundamental properties.

Definition 8.1 (Gaussian (Normal) Distribution)**.** A random variable X follows a Gaussian (or Normal) distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if its probability density function (PDF) is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

If $\mu = 0$ and $\sigma^2 = 1$, it is called a **standard Gaussian**.

In this course, we often work with vectors of Gaussians.

Definition 8.2 (Multivariate Gaussian Distribution)**.** A random vector $\vec{X} \in \mathbb{R}^n$ follows a multivariate Gaussian distribution with mean

vector $\vec{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ (where Σ is positive definite), denoted $\vec{X} \sim N(\vec{\mu}, \Sigma)$, if its PDF is:

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right).$$

If $\vec{X}$ consists of n independent standard Gaussians, then $\vec{\mu} = 0$ and $\Sigma = I_n$ (the identity matrix).

- **Scaling:** If $X \sim N(\mu, \sigma^2)$, then $cX \sim N(c\mu, c^2\sigma^2)$.
- **Sums:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$.

Recall some fundamental properties of Gaussian (Normal) distributions.

- **Scaling:** If $X \sim N(\mu, \sigma^2)$, then $cX \sim N(c\mu, c^2\sigma^2)$.
- **Sums:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$.

A crucial property we will use extensively relates to projections of multivariate Gaussians.

Fact 8.3. Let $G_i \sim N(0,1)$ be independent standard Gaussians, and let $\vec{G} = (G_1, G_2, \ldots, G_n) \in \mathbb{R}^n$. For any fixed vector $x \in \mathbb{R}^n$, the inner product is distributed as:

$$\langle x, G \rangle \sim N\left(0, \sum x_i^2\right) = N(0, \|x\|^2).$$

If $\|x\| = 1$, then $\langle x, G \rangle \sim N(0,1)$.

8.2.1 *Gaussians in Subspaces*

We can also define Gaussian distributions constrained to a subspace.

Definition 8.4 (Gaussian from a Subspace)**.** Suppose V is a subspace of $\mathbb{R}^n$ of dimension $d \leq n$. Pick an orthonormal basis $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d$ for V. We define a **Gaussian from subspace V**, denoted $g \sim N(V)$, as:

$$g = \sum_{i=1}^{d} g_i \vec{v}_i,$$

where $g_i \sim N(0,1)$ are independent.

Fact 8.5. Let $g \sim N(V)$. For any vector $x \in \mathbb{R}^n$:

$$\langle x, g \rangle \sim N(0, \sigma^2)$$

where $\sigma^2 = \|\text{Proj}_V(x)\|^2$. If $\|x\| = 1$, then $\sigma^2 \leq 1$.

8.3 Recap of Martingale Concentration

We recall the Azuma-Hoeffding inequality and its extensions.

Let $Z_1, Z_2, \ldots$ be independent r.v.s, and suppose X_i is a function of $Z_1, \ldots, Z_i$. If the sequence of differences behaves nicely, we have concentration. For example, if $X_i | Z_1, \ldots, Z_{i-1}$ behaves like a Rademacher variable (taking values ± 1 with probability $1/2$), Azuma-Hoeffding gives:

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2T}\right).$$

8.3.1 Gaussian Concentration for Martingales

This concentration extends naturally to Gaussian random variables.

Theorem 8.6 (Gaussian Concentration for Martingales)**.** *Suppose we have a martingale difference sequence such that $X_i | Z_1, \ldots, Z_{i-1}$ is Gaussian with mean 0.*

1. *If the conditional variance is 1, then:*

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2T}\right).$$

2. *More generally, if $X_i | Z_1, \ldots, Z_{i-1} \sim N(0, \sigma_i^2)$, then:*

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2\sum \sigma_i^2}\right).$$

Crucially, these concentration bounds also hold when T is a **stopping time** (not a fixed quantity).

Definition 8.7 (Stopping Time)**.** T is a stopping time with respect to a sequence $X_1, X_2, \ldots$ if the event $\{T = t\}$ depends only on the values of $X_1, \ldots, X_t$.

8.4 Discrepancy Minimization

We now turn to the main application: discrepancy minimization.

Definition 8.8 (Discrepancy)**.** Given a set system $\mathcal{S} = (S_1, S_2, \ldots, S_m)$ where each $S_i \subseteq [n] = \{1, \ldots, n\}$. A 2-coloring of $[n]$ is a map $\chi : [n] \to \{-1, 1\}$. The **discrepancy** of this coloring is

$$\operatorname{disc}(\chi) = \max_{i \in [m]} \left|\sum_{j \in S_i} \chi(j)\right|.$$

We want to find a coloring χ that minimizes the discrepancy (achieves good balance).

8.4.1 Randomized Coloring

Fact 8.9. Consider a simple randomized approach: set $\chi(j) \in \{-1, 1\}$ uniformly and independently at random. For any set S_i, $E\left[\sum_{j \in S_i} \chi(j)\right] = 0$. By Chernoff-Hoeffding bounds:

$$\Pr\left(\left|\sum_{j \in S_i} \chi(j)\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2n}\right).$$

By setting $\lambda = O(\sqrt{n \log m})$ and taking a union bound over all m sets, we find that the discrepancy is $\leq \lambda$ with high probability (w.h.p. $1 - 1/\text{poly}(m)$).

However, we can achieve tighter bounds.

Theorem 8.10 (Spencer's Theorem)**.** *There exists a coloring χ such that*

$$disc(\chi) \leq O\left(\sqrt{n \log(m/n)}\right).$$

If $m = n$, this guarantees $O(\sqrt{n})$ discrepancy. (This is often summarized as "six standard deviations suffice").

Spencer's original proof was non-constructive (using the entropy/pigeonhole principle on an exponentially large family). Bansal (2010) provided the first algorithmic proof using semidefinite programming and rounding. We will present a proof due to Lovett and Meka, which uses Linear Algebra, Gaussians, and Martingales.

8.4.2 Step 1: Relaxation to Partial Colorings

Instead of requiring $\chi : [n] \to \{-1, 1\}$, we consider a "convex" fractional relaxation, allowing $\chi : [n] \to [-1, 1]$.

Bad news: If we only minimize the fractional discrepancy, the problem is trivial: set $\chi(j) = 0$ for all j, achieving "zero fractional discrepancy".

Fix: We require that most variables are "close" to ± 1, allowing only a small fraction of variables to be far from $\{-1, 1\}$.

Lemma 8.11 (Partial Coloring Lemma (Lemma 1))**.** *Let $x_0 = 0$ be the starting point in $[-1, 1]^n$. We can find $x \in [-1, 1]^n$ such that:*

1. $\left|\sum_{j \in S_i} x_j\right| \leq \sqrt{|S_i|} \cdot \Delta + 1/poly(n)$ *for all i.*
2. $\#\{j \text{ s.t. } x_j \notin \{-1, 1\}\} \leq n/2$.

Here $\Delta = c\sqrt{\log(m/n)}$ for some constant c.

We will actually prove a more general statement involving arbitrary vectors, which implies Lemma 8.11.

Lemma 8.12 (Generalized Partial Coloring Lemma (Lemma 2))**.** *Given any vectors $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, any starting point $x_0 \in [-1,1]^n$, and a small $\delta > 0$ (e.g., $1/poly(n)$). We can find $x \in [-1,1]^n$ such that:*

1. $|\langle a_i, x - x_0\rangle| \leq \|a_i\|_2 \cdot \Delta$ *for all i.*
2. $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq n/2$.

Here $\Delta = c\sqrt{\log(m/n)}$.

To see that Lemma 8.12 implies Lemma 8.11, set a_i to be the indicator vector of set S_i (so $\|a_i\|_2 = \sqrt{|S_i|}$) and set $x_0 = 0$. We then take the solution given by Lemma 8.12 and round the variables close to ± 1 to exactly ± 1.

8.4.3 *From Partial to Total Coloring*

Before proving Lemma 8.12, let's see how Lemma 8.11 implies Spencer's Theorem. We use an iterative approach.

Start at $X_0 = 0$. Apply Lemma 8.11. This yields a set J of $\geq n/2$ variables colored $\{\pm 1\}$. The remaining $\leq n/2$ variables are fractional.

Freeze the variables in J. Consider the remaining variables $[n] \setminus J$. Start where the previous run stopped (using that configuration as the new X_0) and run Lemma 8.11 again on this smaller instance of size $\leq n/2$.

This finds a new set J' of $\geq \frac{1}{2}|[n] \setminus J|$ variables at ± 1. Repeat.

The net discrepancy may add up over the iterations. The total discrepancy is bounded by:

$$\begin{aligned}\chi(\mathcal{S}) &\leq c\sqrt{n\log(m/n)} + c\sqrt{\frac{n}{2}\log(m/(n/2))} + c\sqrt{\frac{n}{4}\log(m/(n/4))} + \ldots \\ &= c \cdot \sum_{i \geq 0} \sqrt{\frac{n}{2^i}\log\left(\frac{2^i m}{n}\right)} \\ &= O(\sqrt{n\log(m/n)}).\end{aligned}$$

The sum converges, dominated by the first term. This shows how to get Spencer's Theorem from Lemma 1 (and hence from Lemma 2).

8.5 *Proof of Lemma 8.12: Gaussian Random Walks*

How to prove Lemma 8.12? We use a beautiful algorithm utilizing ideas from:

- Random Walks
- Gaussians
- Martingales

Recall the goal (Lemma 8.12). WLOG, assume a_i are unit vectors, so we want $|\langle a_i, x - x_0 \rangle| \leq \Delta$. For convenience, we will prove that $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq 7n/10$ (instead of $\leq n/2$).

8.5.1 *The Algorithm Idea*

The idea is to start at x_0 and take tiny Gaussian steps. ($X^{t+1} = X^t +$ gaussian).

If we just do this, it is not great. We need to maintain constraints:

1. Coordinate bounds: $x_j \in [-1, 1]$ for all j.

2. Discrepancy bounds: $\langle a_i, x - x_0 \rangle \in [-\Delta, \Delta]$ for all i.

The Key Idea: If we get close to violating some constraint, we "freeze" the solution, forcing subsequent steps to lie in a subspace orthogonal to that constraint.

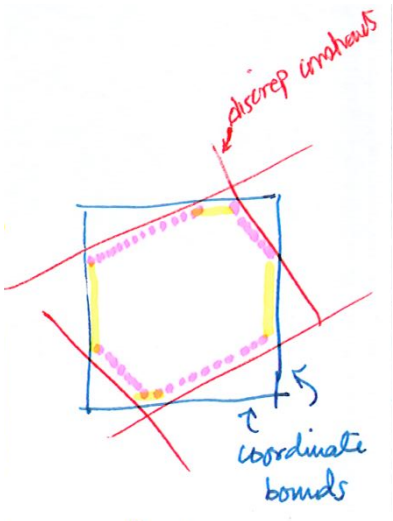

Figure 8.1: Feasible Region for the Random Walk.

We define when a variable or a constraint is close to being violated.

Definition 8.13. A variable j is **frozen** if $|x_j| > 1 - \delta$. A constraint a_i is **dangerous** if $|\langle a_i, x - x_0 \rangle| > \Delta - \delta$.

Intuition: The variable bounds are often much closer to the origin than the discrepancy bounds. So we expect the random walk to hit a variable bound earlier, meaning we will freeze many more variables than constraints become dangerous.

8.5.2 *The Algorithm Details*

At time t, we have a solution x^t. Let $F^t = \{j : |x_j^t| \geq 1 - \delta\}$ be the set of frozen variables. Let $D^t = \{i : |\langle a_i, x^t - x^0\rangle| \geq \Delta - \delta\}$ be the set of dangerous constraints.

We define the subspace V_t where the next step must lie.

Definition 8.14. Let V_t be the subspace orthogonal to all frozen variables and all dangerous constraints.

$$V_t = \left(\text{span}\left(\{e_j : j \in F^t\} \cup \{a_i : i \in D^t\}\right)\right)^{\perp}.$$

The algorithm proceeds as follows. We set the step size $\epsilon \leq \frac{\delta}{10\sqrt{\log(mn)}}$.

Algorithm 3: Lovett-Meka Algorithm

2.1 **while** $\dim(V_t) \geq n/2$ **do**
2.2 (This means we have few frozen/dangerous constraints/variables.)
2.3 Pick $g_t \sim N(V_t)$ (a Gaussian from that subspace).
2.4 $x^{t+1} \leftarrow x^t + \epsilon \cdot g_t$.

The algorithm stops when $\dim(V_t) < n/2$. This means we have at least $n/2$ frozen variables or dangerous constraints in total (or more precisely, the dimension spanned by them is $> n/2$). We need to show that most of these are frozen variables.

8.5.3 *Analysis of the Algorithm*

What could go wrong with the algorithm?

1. The solution x^t "jumps" outside the feasible region ($[-1, 1]^n$ or the discrepancy bounds).
2. The algorithm stops (when $\dim(V_t) < n/2$), but very few variables are frozen (mostly dangerous constraints).

Let's address these concerns.

1. Staying Feasible. We know that x^{t-1} was "good". For non-frozen variables, $|x_j^{t-1}| \leq 1 - \delta$. For non-dangerous constraints, $|\langle a_i, x^{t-1}\rangle| \leq \Delta - \delta$.

For x^t to go outside the feasible region, the step must be large: $|\epsilon g^t| \geq \delta$.

We analyze the probability of a large Gaussian step. Since the components of g^t have variance ≤ 1:

$$\Pr(|\epsilon g^t| \geq \delta) \leq 2\exp\left(-\frac{(\delta/\epsilon)^2}{2}\right).$$

We chose ϵ such that $\delta/\epsilon \geq \sqrt{\log(mn)}$.

$$\Pr(|\epsilon g^t| \geq \delta) \leq 2\exp\left(-\frac{100\log(mn)}{2}\right) = \frac{2}{(mn)^{50}}.$$

We can take a union bound over all $m+n$ constraints and over all time steps, provided the total number of steps T is less than $(mn)^{49}$. We will actually prove that $T = O(1/\epsilon^2) = O(\text{poly}(m,n))$ many steps.

2. How many steps? We analyze the expected progress of the algorithm using the ℓ_2 norm.

Fact 8.15. $E\left[\|x^T - x^0\|^2\right] = \sum_{t<T} \epsilon^2 \cdot E[\dim(V_t)]$.

Proof.

$$\begin{aligned} E\left[\|x^{t+1} - x^0\|^2\right] &= E\left[\|x^t - x^0 + \epsilon g_t\|^2\right] \\ &= E\left[\|x^t - x^0\|^2\right] + 2\epsilon E\left[\langle g_t, x^t - x^0\rangle\right] + \epsilon^2 E\left[\|g_t\|^2\right]. \end{aligned}$$

The middle term is o because g_t is independent of x^t (conditioned on V_t) and $E[g_t] = 0$ (by symmetry). The last term $E\left[\|g_t\|^2 \mid V_t\right]$ is the dimension of the subspace V_t. So, $E\left[\|x^{t+1} - x^0\|^2\right] = E\left[\|x^t - x^0\|^2\right] + \epsilon^2 \cdot E[\dim(V_t)]$. The result follows by induction. □

Since the algorithm runs as long as $\dim(V_t) \geq n/2$, we have:

$$E\left[\|x^T - x^0\|^2\right] \geq \sum_t \epsilon^2 \cdot \frac{n}{2} = T \cdot \frac{n}{2}\epsilon^2.$$

On the other hand, since x^T must remain in $[-1,1]^n$, we must have $\|x^T - x^0\|^2 \leq O(n)$.

So, $T\frac{n}{2}\epsilon^2 \leq O(n)$. This implies $T \leq O(1/\epsilon^2)$. The algorithm is very likely to stop after $O(1/\epsilon^2) = O(\text{poly}(n,m))$ steps.

3. What happens at the stopping time? We need to analyze how many frozen vs dangerous constraints we have. We want to bound the number of dangerous constraints.

Let's see what the probability is that a specific constraint i (say a_i) becomes dangerous. Let $Y_t = \langle a_i, x^t - x^0\rangle$. We want $\Pr(|Y_T| > \Delta - \delta)$.

Y_T is a martingale (it is the sum of the noise contributions $\epsilon\langle a_i, g_t\rangle$). Let $Z_t = \epsilon\langle a_i, g_t\rangle$. Z_t is Gaussian $N(0, \sigma_t^2)$ where $\sigma_t^2 \leq \epsilon^2$ (since a_i is a unit vector and g_t is a Gaussian in a subspace).

We use the Gaussian concentration for martingales, noting that T is a stopping time.

$$\begin{aligned} \Pr(|Y_T| > \Delta - \delta) &\leq 2\exp\left(-\frac{(\Delta-\delta)^2}{2\sum E[\sigma_t^2]}\right) \\ &\leq 2\exp\left(-\frac{(\Delta-\delta)^2}{2T\epsilon^2}\right). \end{aligned}$$

We know $T\epsilon^2 = O(1)$. We set $\Delta = O(\sqrt{\log(m/n)})$.

$$\Pr(\text{constraint } i \text{ dangerous}) \leq 2\exp(-O(\Delta^2)) = 2\exp(-O(\log(m/n))) \leq \frac{n}{10m}.$$

(By choosing the constant c in Δ large enough).

Now we can calculate the expected number of dangerous constraints:

$$E[\#\text{dangerous constraints}] = \sum_{i=1}^{m} \Pr(\text{constraint } i \text{ dangerous}) \leq m \cdot \frac{n}{10m} = n/10.$$

By Markov's inequality,

$$\Pr(\#\text{dangerous} \geq n/5) \leq 1/2.$$

When the algorithm stops, we have $\dim(V_t) < n/2$. This implies that the span of frozen variables and dangerous constraints has dimension $> n/2$. With probability $\geq 1/2$, we have #Dangerous $< n/5$. In this case, the number of frozen variables must be large enough to account for the remaining dimension (e.g., #Frozen $\geq n/2 - n/5 = 3n/10$).

This successfully proves Lemma 8.12 (except that we proved $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq 7n/10$ instead of $\leq n/2$.)).

8.6 To wrap up

Today we saw Gaussian RVs and their use for:

- Dimension reduction (for distance preservation).
- Discrepancy minimization.

Along the way, we needed:

- Concentration bounds for sums of squares of (independent) Gaussians (Chi-squared distribution).
- Concentration bounds for Gaussians (but using martingale techniques).
- Random walks with Gaussians.

These ideas are simple but very powerful! The dimension reduction techniques (approximate) are useful in various surprising contexts, such as Compressive Sensing (also the "single-pixel camera").