Thoughts towards high probability bounds for gradient descent

1 The Toy Problem

Recursive formula A. Define the following stochastic process.

$$\begin{aligned} B_1 &= 1 \\ B_{i+1} &= \frac{i-1}{i} B_i + \frac{1}{2}\sqrt{\frac{B_i}{i}} \cdot U_i + \frac{1}{i}, \end{aligned}$$

where $U_1, U_2, ...$ are i.i.d., uniform on $\{-1, +1\}$. (The value 1/2 here is not critical, and could be replaced by any constant less than 1.)

Recursive formula B. Another way to write that is

$$\begin{aligned} B_1 &= 1 \\ B_{i+1} - B_i &= \frac{1 - B_i}{i} + \frac{1}{2}\sqrt{\frac{B_i}{i}} \cdot U_i. \end{aligned}$$

2 The Conjectures

The cleanest conjecture that seems to capture the heart of the problem is as follows.

Conjecture 2.1. There exists a constant c such that for any n, $\Pr\left[\, B_n \geq c \log(1/\delta) \,\right] \leq \delta$.

The actual conjecture that we want to prove is as follows.

Conjecture 2.2. There exists a constant c such that for any n, $\Pr\left[\sum_{i=n/2}^{n} \frac{B_i}{i} \geq c \log(1/\delta)) \right] \leq \delta$.

I believe this second conjecture should be relatively straightforward once the first conjecture is proven. A stronger conjecture is the following.

Conjecture 2.3. There exists a constant c such that for any n, $\Pr\left[\, \frac{1}{n} \sum_{i=1}^{n} B_i \geq c \log(1/\delta)) \,\right] \leq \delta$.

This obviously implies Conjecture 2.2.

2.1 What is too strong?

I would guess that the following is not true.

Presumably False 2.4. $\exists c$ such that $\forall n$, $\Pr\left[\, \exists i \leq n \text{ s.t. } B_i \geq c \log(1/\delta)) \,\right] \leq \delta$.

So the challenge is to find an argument that is not so strong that it would imply this, but it would imply Conjecture 2.1 (and Conjecture 2.2).

2.2 Connection to stochastic gradient descent

In the paper of Rakhlin et al., they want to prove that $\sum_{i=n/2}^{n} \|w_i - w^*\|^2 = O(\log(1/\delta))$ with probability at least $1-\delta$. This basically amounts to defining $B_i = i\,\|w_i - w^*\|^2$ and proving Conjecture 2.2.

3 Some basic results

Claim 3.1. $\mathrm{E}[\,B_n\,] = 1$ for all n.

Proof. By induction, the case $n=1$ being trivial. Next, using that U_i is independent of B_i,

$$\begin{aligned}
\mathrm{E}[\,B_{i+1} \mid B_i\,] &= \frac{i-1}{i}\,\mathrm{E}[\,B_i \mid B_i\,] + \mathrm{E}\left[\sqrt{\frac{B_i}{i}}\cdot U_i \;\Big|\; B_i\right] + \frac{1}{i} \\
&= \frac{i-1}{i} B_i + \sqrt{\frac{B_i}{i}} \cdot \underbrace{\mathrm{E}[\,U_i \mid B_i\,]}_{=0} + \frac{1}{i} \\
&= \frac{i-1}{i} B_i + \frac{1}{i}
\end{aligned}$$

Thus, taking the expectation and using the inductive hypothesis,

$$\mathrm{E}[\,B_{i+1}\,] = \frac{i-1}{i}\underbrace{\mathrm{E}[\,B_i\,]}_{=1} + \frac{1}{i} = \frac{i-1}{i} + \frac{1}{i} = 1.$$

■

Non-recursive formula. Let us now derive a non-recursive formula for B_n.

Claim 3.2. For all $n \geq 1$,

$$B_{n+1} = 1 + \frac{1}{2n} \sum_{i \leq n} \sqrt{iB_i} \cdot U_i.$$

Proof. By induction. For $n=1$, this is immediate from the recursive definition. For $n>1$, we have

$$\begin{aligned}
B_{n+1} &= \frac{n-1}{n} B_n + \frac{1}{2}\sqrt{\frac{B_n}{n}} \cdot U_n + \frac{1}{n} \\
&= \frac{n-1}{n}\left(1 + \frac{1}{2(n-1)} \sum_{i \leq n-1} \sqrt{iB_i} \cdot U_i\right) + \frac{1}{2}\sqrt{\frac{B_n}{n}} \cdot U_n + \frac{1}{n} \qquad \text{(inductive hypothesis)} \\
&= \frac{n-1}{n} + \frac{1}{2n} \sum_{i \leq n-1} \sqrt{iB_i} \cdot U_i + \frac{1}{2n}\sqrt{nB_n} \cdot U_n + \frac{1}{n}
\end{aligned}$$

$$= 1 + \frac{1}{2n} \sum_{i \le n} \sqrt{iB_i} \cdot U_i,$$

which completes the proof. ■

Claim 3.3. For all $n \ge 1$, $\Pr[\, B_n > 0 \,] = 1$.

Proof. By induction, the cases $n = 1$ and $n = 2$ being obvious from the definition. So let $n \ge 2$, assume $\Pr[\, B_n > 0 \,] = 1$, and let us prove the claim for $n + 1$.

Let $x = \sqrt{B_n}$ (which is well-defined since B_n is positive). From the recursive definition of B_{n+1}, we have

$$B_{n+1} \ \ge \ \frac{n-1}{n} \cdot x^2 - \frac{1}{2\sqrt{n}} \cdot x + \frac{1}{n}.$$

This is a quadratic equation in x, and therefore it has no roots iff its discriminant is negative. The discriminant is

$$\frac{1}{4n} - 4\frac{n-1}{n^2} \ = \ \frac{1}{4n}\left(1 - 16\frac{n-1}{n}\right),$$

which is obviously negative for all $n \ge 2$. Therefore B_{n+1} is positive. ■

4 Heuristic justification for conjectures

4.1 Similarity to simple ±1 random walk

As above, let $U_1, U_2, ...$ be i.i.d., uniform on ± 1. Let $S_n = \sum_{i=1}^n U_i$. Let $B_n = S_n / \sqrt{n}$. Then $\mathrm{E}[\, B_n \,] = 0$ for all n. Azuma's inequality implies that

$$\Pr\left[\, \exists i \le n \text{ s.t. } S_i \ge c\sqrt{n \log(1/\delta)} \,\right] \le \delta.$$

Thus, a consequence is

$$\Pr\left[\, B_n \ge c\sqrt{\log(1/\delta)} \,\right] \le \delta,$$

which is the analog of Conjecture 2.1. Furthermore, another consequence is

$$\Pr\left[\, \forall i \in \{n/2, ..., n\}\, B_i \ge 2c\sqrt{\log(1/\delta)} \,\right] \ge 1 - \delta.$$

Thus

$$\Pr\left[\, \sum_{i=n/2}^{n} \frac{B_i}{i} \ge 2c\sqrt{\log(1/\delta)} \,\right] \ge 1 - \delta,$$

which is the analog of Conjecture 2.2.

However, (I think) it is *not* true that

$$\Pr\left[\, \exists i \le n \text{ s.t. } S_i \ge c\sqrt{i \log(1/\delta)} \,\right] \le \delta,$$

as (I think) this would contradict the law of the iterated logarithm. Thus, it is *not* true that

$$\Pr\left[\, \exists i \le n \text{ s.t. } B_i \ge c\sqrt{\log(1/\delta)} \,\right] \le \delta.$$

This is the analog of Presumably False 2.4.

4.2 Drift comparable to variance

I notice one interesting thing about Recursive Formula B. Let's suppose that $B_t \gg 1$, so that we have

$$B_{i+1} - B_i \;\approx\; c\sqrt{\frac{B_i}{i}} \cdot U_i \;-\; \frac{B_i}{i}.$$

(for $c = 1/2$, say). We can think of the $-B_i/i$ as a "drift term" that makes B_{i+1} decrease towards its expectation (which is 1). But the variance associated with the random increment is

$$\left(c \cdot \sqrt{B_i/i}\right)^2 \;=\; c^2 B_i/i.$$

So the total drift up to time T is

$$- \sum_{i \leq T \text{ s.t. } B_i \text{ is large}} B_i/i$$

And the total variance from the randomness up to time T is

$$\sum_{i \leq T} \left(c \cdot \sqrt{B_i/i}\right)^2 \;=\; c^2 \sum_{i \leq T} B_i/i.$$

Intuitively, any positive amount due to the randomness should be canceled out by the negative amount due to the drift. But I'm not sure how to make this precise, because the drift mainly contributes its negative amount when $B_i \gg 1$, whereas the random increments occur all the time...