Speaker: Matus Telgarsky (UIUC)

The goal of this talk is to prove that stochastic gradient descent (sgd) converges with high probability to something meaningful in the task of logistic regression without any constraints or regularization, in a regime where iterate norms may grow without bound. In this regime, it is shown that the dual iterates—and hence the logistic probability outputs—always converge to the dual optimum. If a certain margin property of the distribution is large, then the dual rate is $O(t^{-1/4})$, and the (primal) risk also converges at a rate $O(t^{-1/2})$. It is also shown that a modified algorithm—not sgd, but still performing one pass of the data and using $O(d^2)$ storage—similarly always converges, and under the same margin property has dual rate $O(t^{-1/2})$ and primal rate $O(t^{-1})$. By contrast, existing analyses in these regimes do not appear to be better than $O(1)$.

There's no boundedness in the parameter space. I'm not talking about boundedness in the covariates.

1 Setting and goal

The problem is open.

Why only the logistic loss?

Find $w \in \mathbb{R}^d$ to minimize

$$R(w) = \int \ell(\langle w, -xy\rangle)\, d\mu(x,y)$$

I'm talking about population risk.

I'm not using projection, and have a strong condition on the stochasticity in the SGD,

$$w_0 = 0, \quad w_s = w_{s-1} - \eta g_s, \quad \forall s > 0$$

where $g_s = -\ell'(\langle w_{s-1}, -x_s y_s\rangle) x_s y_s$ for iid $(x_s, y_s) \sim \mu$. I need this! We have to weave martingale throubh sgd recursion analysis.

Goal: with probability $\geq 1 - de$

$$R(\widehat{w}_t) - \inf_w R(w) = \widetilde{O}\left(\frac{(\text{dist const}), \ln\left(\frac{1}{\delta}\right)}{t}\right)$$

Note we're comparing against the infimum, not minimum in bounded set.

Prior work have dependence on norms

$$R(\widehat{w}_t) - \inf_w R(w) = R(u) - \inf_w R(w) + \widetilde{O}\left(\frac{\max_{s<t} \|w_s - u\|_2 \ln\left(\frac{1}{\delta}\right)}{\sqrt{t}}\right).$$

Alternatively, use strong convexity, but the strong convexity constant decays exponentially with the ball you're searching over.

If perfect separator, infimum loss is 0 but $R(w) > 0$ for all w. The infimum is not attained! This phenomenon does not happen with squared loss.

We get $\widetilde{O}(1)$ without analysis of w_s, u. $\widetilde{O}\left(\frac{1}{t}\right)$ requires martingale analysis.

This should be believable.

2 Motivation

Who cares?

- SGD is often minimally constrained/regularized in practice.
- "Understanding NN requires rethinking generalization" says: This phenomenon is qualitatively unaffected by explicit regularization.
- I wondered if it was possible to analyze things without boundedness conditions.

Consider experiment: minimize $\widehat{R}_n(w) + \frac{1}{2n^p} \|w\|^2$ for $p \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, \infty\}$. We have a finite training set, minimize using L-BFGS.

For the risk bound go to 0 with number of samples (analyzing with Rademacher complexity), you need to use one of first three choices.

As p increases past threshold, $p \to \infty$, for most of the datasets, classification error doesn't go up!

3 Risk structure

Study risk and the structure that this imposes.

There is an ancient analysis when there are margins. We will prove the general case is a combination of the two cases (bounded vs. margin).

What properties of the measure μ induce bounded sublevel sets?

This doesn't require anything about convex loss except it's 0.

Under the structure where every direction makes classification errors (convex hulls of +/- examples overlap?), we get bounded sublevel sets and global optimum. Convex losses don't like it when you make mistakes. It will not allow you to be too big.

Second case: some direction makes no error. This is the classical setting of margins. Covering number is based on the amount of margin you have. We have a lot of wiggle room for the direction.

Okay if arbitrarily small mass close to margin: for all $\varepsilon > 0$, exists $\gamma(\varepsilon) > 0$, $\overline{u}$, $\mathbb{P}[\langle \overline{u}, xy \rangle < \gamma(\varepsilon)] \leq \varepsilon$.

There is another case, the mixed regime. Earlier we had two densities that looked continuous. Now consider if we have positive probability on 2 lines.

(Picture.) "Optimal" vector could be a finite point plus an infinite vector in the orthogonal directions.

$$\lim_{c \to \infty} R(\overline{v} + c \cdot \overline{u}) = \inf_w R(w).$$

Theorem 3.1 (Structure theorem)**.** *There exist unique $S, S^\perp$ such that*

- *There exists unique $\overline{v} \in S$,*

$$R_S(\overline{v}) = \inf_w R_S(w) = \inf_w R(w).$$

- *Given* $(w_i)_{i\geq 1}$,

$$\Pi_S(w_i) \to \overline{v} \tag{1}$$
$$R_{S^c}(w_i) \to 0 \tag{2}$$
$$R_{S^c}(\Pi_{S^\perp}(w_i)) \to 0. \tag{3}$$

(I have to talk about minimizing sequences because norms go to ∞.) The last line is loss-sensitive; other things can be proved for losses that are strictly convex and asymptote to 0 on one side. Effectively like exponential loss as $\to -\infty$.

"Logistic loss is good." It's the only loss that many people use for classification.

Essence of proof: dissect the dual. The dual optimum always exists!

4 Theorem

I give the $O\left(\frac{1}{\sqrt{t}}\right)$ version. With prob $\geq 1-\delta$,

$$R(\widehat{w}_t) - \inf_w R(w) = \widetilde{O}\left(\frac{\ln\left(\frac{1}{\delta}\right)^2}{\sqrt{t}}(1 + t\varepsilon^2 + \frac{1}{\gamma(\varepsilon)^4})\right).$$

$\gamma(\varepsilon)$ is such that there exists unit vector $\overline{u} \in S^\perp$ s.t. $\mathbb{P}[x \in S^\perp \wedge \langle \overline{u}, xy\rangle < \gamma(\varepsilon)] \leq \varepsilon$. This can blow up, $t\varepsilon^2 + \gamma(\varepsilon)^{-1} \geq \sqrt{t}$. (Ex. If the distribution grows towards the boundary.)

Cheat: for the dual optimum,

$$\int |\widehat{q}_t - \overline{q}|\, d\mu = \widetilde{O}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{t^{\frac{1}{4}}}(1 + t^{\frac{1}{4}}\sqrt{\varepsilon} + \frac{1}{\gamma(\varepsilon)})\right).$$

- I can make ε small and wait for t to be large. The dual is the probability prediction—the probability predictions do converge. Here $\widehat{q}_t(x,y) = t^{-1}\sum_{s<t}\ell'(\langle w_s, -xy\rangle)$.
- This means 0-1 loss converges quickly in general; we don't know about the risk.
- The $\widetilde{O}$ is hiding $\operatorname{poly}\log(t), \operatorname{poly}\log(d)$. The dependence depends on how big sublevel sets are.

5 Proof

The proof strategy: separate analysis on S and $S^\perp$.

- Over S: $\|w_t - \overline{v}\|$.
- Over $S^\perp$: $\left\|\frac{w_t}{\|w_t\|} - \overline{u}\right\|$.

The problem decouples. Look instead at

- $|\Pi_S(w_t) - \overline{v}|$.
- $\min \left\| \frac{\Pi_{S^\perp}(w_t)}{\|\Pi_{S^\perp}(w_t)\|} - \overline{u} \right\|$ or $\langle w_t, \overline{u} \rangle$.

Over S do usual analysis. Over $S^\perp$ do perceptron analysis. Cross terms are small and easy to control. Proving the decomposition into subspaces is the hard part. The convergence loss hadn't been done for logistic loss, just for hinge.

1. Why cross terms vanish: $\|\Pi_S(w_j)\| \leq B$. Take usual smoothness proof: Write smoothness upper bound, do algebra, gradient descent converges to something with small gradient. One step of this tells us you stay within sublevel set.

   Let $v_t = \Pi_S(w_j)$. We control the potential function

   $$\|v_{t+1} - v\|^2 = \|v_t - \overline{v}\|_2^2 + 2\eta \langle g_{t+1}, \overline{v} - v_t \rangle + \eta^2 \|\Pi_S(y_{t+1})\|^2 \tag{4}$$

   the first term should worry you. The first term is

   $$= \langle g_{t+1} - \nabla R(w_t), \overline{v} - v_t \rangle + \underbrace{\langle R(w_t), \overline{v} - v_t \rangle}_{\langle \nabla R_S(v_t), \overline{v} - v_t \rangle + \left\langle \nabla R_{S^\perp}(w_t), \overline{v} - v_t \right\rangle} \tag{5}$$

   These terms in the underbrace are respectively $\leq R_S(\overline{v}) - R_S(v_t)$, population version of mistake bound $2B \sum_{j<t} \int_{S^\perp} \ell'(\langle w_S, -xy \rangle)\, d\mu(x,y)$. Control using perceptron proof.

2. $$\inf_w \int \ell(\langle w, xy \rangle)\, d\mu(x,y) = \sum \left\{ \int -\ell^*(q(x,y))\, d\mu(x,y) : q \in L^1(\mu), \int xyq(x,y) = 0 \right\}, \tag{6}$$

   Interpreting the condition: it says for all w, $\int \langle w, xy \rangle\, q(x,y) = 0$, a measure which decorrelates everything. Equally supported in correct/wrong predictions. Why you get boundedness over set. (See Rockefeller paper)

   $$\int_{\langle w, xy \rangle > 0} \langle w, xy \rangle\, q(x,y) = \int_{\langle w, xy \rangle < 0} \langle w, -xy \rangle\, q(x,y).$$