COS-511: Learning Theory **Spring 2017**

Lecture 13: Sample Complexity of Linear Classes & SGD

Lecturer: Roi Livni

13.1 The Sample Complexity of Linear Classes

We now apply the tools we developed so far to bound the complexity of learning linear classes. As an application we will consider learning surrogate loss functions for two cases: Learning ℓ_2 regulerized classifiers over the unit ball, and learning ℓ_1 regulerized classifiers over the cube. Recall that our objective is to bound the performance of an empirical risk minimizer. Namely, given a learning problem $(\mathcal{X}, \mathcal{H}, \ell)$ we denote

$$\mathcal{L}_D(f) = \mathbf{E}_{(\mathbf{x},y)\sim D}(\ell(f(\mathbf{x}), y)) \quad \mathcal{L}_S = \frac{1}{m}\sum_{i=1}^{m} \ell(f(\mathbf{x}^{(i)}), y_i)$$

Further, denote

$$\ell \circ \mathcal{H} = \{\ell(f(\mathbf{x}^{(i)}), y_i) : f \in \mathcal{H}\}$$

Then so far we've shown that w.p $1-\delta$: for each $f_\ell \in \ell \circ \mathcal{H}$:

$$\mathcal{L}_D(f_\ell) \leq \mathcal{L}_S(f_\ell) + \mathfrak{R}_m(\ell \circ \mathcal{H}) + O(\sqrt{\frac{\log 1/\delta}{m}})$$

Note that if $f_\ell(\mathbf{x}, y) = \ell(f(\mathbf{x}, y)$ and $(\mathbf{x}, y)$ are sampled from D then $\mathcal{L}_D(f_\ell)$ measures the expected loss of f and similarly $\mathcal{L}_S$ w.r.t the empirical loss. Our aim now is to bound $\mathfrak{R}_m(\ell \circ \mathcal{H})$: We will consider the Rademacher complexity of special classes of problems:

13.1.0.1 Learning ℓ_2 regulerized classifiers

We consider learning problem $(\mathcal{X}, \mathcal{H}, \ell)$, where ℓ is convex and L-Lipschitz[1], $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$ and $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\}$, here we identify each $\mathbf{w} \in \mathcal{H}$ with the operation $\mathbf{x} \to \mathbf{w} \cdot \mathbf{x}$.

Lemma 13.1. *Let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$, be the unit ball of some Hilbert–space: Set $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B,\ \mathbf{w} : \mathbf{x} \to \mathbf{w} \cdot \mathbf{x}\}$ and ℓ be a L-Lipschitz loss function then, let $\mathbf{w}_S \in \mathcal{H}$ be a function such that:*

$$L_S(\mathbf{w}_S) \leq \min_{\mathbf{w}^* \in \mathcal{H}} L_S(\mathbf{w}) + O(\frac{LB}{\sqrt{m}})$$

then for every $\mathbf{w}^*$ w.p. $1-\delta$:

$$\mathcal{L}_D(\mathbf{w}) \leq \mathcal{L}_D(\mathbf{w}^*) + O\left(LB\sqrt{\frac{\log 1/\delta}{m}}\right).$$

[1]Recall that a function f is L-Lipschitz if $|f(a) - f(b)| \leq L$ an alternative defintion for smooth functions is $\|\nabla f\| \leq L$

Proof. Racall that $\mathfrak{R}(B \cdot \mathcal{F}) = |B|\mathfrak{R}(\mathcal{F})$ (see fact 12.1). Hence applying Lem. 12.3 we get that $\mathfrak{R}(\mathcal{H}) = B\mathfrak{R}(\mathcal{H}_1)$ where $\mathcal{H}_1$ is the class of 1-norm bounded linear classifiers.

Applying Lem. 12.4 with $\phi_{(\mathbf{x},y)}(\mathbf{w}) = \ell(\mathbf{w} \cdot \mathbf{x}, y)$ we obtain that

$$\mathfrak{R}_m(\ell \circ \mathcal{H}) \leq \frac{LB}{\sqrt{m}}$$

As a corollary, w.p $1-\delta$ over a sample $S = (\{\mathbf{x}^{(i)}, y_i\})$ drawn IID, we have for every $\mathbf{w} \in \mathcal{H}$:

$$\mathcal{L}_D(\mathbf{w}) \leq \mathcal{L}_S(\mathbf{w}) + O\left(LB\sqrt{\frac{\log 1/\delta}{m}}\right).$$

In particular for $\mathbf{w}_S$, since $\mathcal{L}_S(\mathbf{w}_S) \leq \mathcal{L}_S(\mathbf{w}^*) + O(LB\sqrt{\frac{\log 1\delta}{m}})$:

$$\mathcal{L}_D(\mathbf{w}) \leq \mathcal{L}_S(\mathbf{w}^*) + O\left(LB\sqrt{\frac{\log 1/\delta}{m}}\right).$$

□

Using standard concentration bound for $\mathcal{L}_S(\mathbf{w}^*)$ we obtain the desried result.

13.1.0.2 Learning ℓ_1 regulerized classifiers

Next, we consider learning problem $(\mathcal{X}, \mathcal{H}, \ell)$, where ℓ is again convex and L-Lipschitz, $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \leq 1\}$ and $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_1 \leq B\}$. We obtain a similar result using Masart Lemma:

Lemma 13.2. *Let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \leq 1\}$, be the unit cube in $\mathbb{R}^d$: Set $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_1 \leq B,\ \mathbf{w} : \mathbf{x} \to \mathbf{w} \cdot \mathbf{x}\}$ and ℓ be a L-Lipschitz loss function then, let $\mathbf{w}_S \in \mathcal{H}$ be a function such that:*

$$L_S(\mathbf{w}_S) \leq \min_{\mathbf{w} \in \mathcal{H}} L_S(\mathbf{w}) + O\left(L\frac{B\log\frac{d}{\delta}}{\sqrt{m}}\right)$$

then for every $\mathbf{w}^*$ w.p. $1-\delta$:

$$\mathcal{L}_D(\mathbf{w}_S) \leq \mathcal{L}_S(\mathbf{w}^*) + O\left(LB\sqrt{\frac{\log\frac{d}{\delta}}{m}}\right).$$

Proof. Let $A = \{\pm\mathbf{e}_i\}_{i=1}^d$ be the standard basis vector in $\mathbb{R}^d$, then note that

$$\mathcal{H} = \text{conv}\{\pm\mathbf{e}_i\},$$

thus we obtain for $\phi_{(\mathbf{x},y)}(\mathbf{w}) = \ell(\mathbf{w} \cdot \mathbf{x}, y)$

$$\mathfrak{R}_m(\ell \circ \mathcal{H}) \leq L\mathfrak{R}_m(\mathcal{F}) = L\mathfrak{R}_m(A)$$

Where first equality is by Lem. 12.4 and second equality is due to property 2 in fact 12.1. Finally by Masart's Lemma (Lem. 12.2), and recaling, we have that;

$$\mathfrak{R}_m(A) \leq B\sqrt{\frac{\log d}{m}}.$$

As a corollary, w.p $1-\delta$ over a sample $S = (\{\mathbf{x}^{(i)}, y_i\})$ drawn IID, we have for every $\mathbf{w} \in \mathcal{H}$:

$$\mathcal{L}_D(\mathbf{w}) \leq \mathcal{L}_S(\mathbf{w}) + O\left(LB\sqrt{\frac{\log d/\delta}{m}}\right).$$

In particular for $\mathbf{w}_S$, since $\mathcal{L}_S(\mathbf{w}_S) \leq \mathcal{L}_S(\mathbf{w}^*) + O(LB\sqrt{\frac{\log \frac{d}{\delta}}{m}})$:

$$\mathcal{L}_D(\mathbf{w}) \leq \mathcal{L}_S(\mathbf{w}^*) + O\left(LB\sqrt{\frac{\log \frac{d}{\delta}}{m}}\right).$$

Again using standard concentration bound for $\mathcal{L}_S(\mathbf{w}^*)$ we obtain the desried result. □

13.2 Stochastic Gradient Descent

In the previous section we devloped tools to analyse the sample complexity for certain learning problems of the form $(\mathcal{X}, \mathcal{H}, \ell)$ where ℓ is Lipschitz and convex. We next address the issue of computation complexity. The subject of Convex Optimization is vast and has seen alot of research, in particular in the context of Machine Learning. Here we will analyse the Stochastic Gradient Descent algorithm, which is perhaps the most studied algorithm in stochastic optimization, and is quite general.

13.2.1 Problem Setup

In stochastic optimizatoin, the optimizer attempts to minimize a convex (Lipschitz) function f over a convex domain $\mathcal{K}$ of diameter D i.e.:

$$D = \max_{\mathbf{x},\mathbf{y}\in\mathcal{K}} \|\mathbf{x} - \mathbf{y}\|$$

The key algorithmic assumption is that the optimizer has access to noisy gradient oracle, defined by:

$$\mathcal{O}(\mathbf{x}) = \hat{\nabla}_{\mathbf{x}}, \qquad \mathbb{E}\left[\hat{\nabla}_{\mathbf{x}}\right] = \nabla f(\mathbf{x}), \quad \mathbb{E}\left[\|\hat{\nabla}_{\mathbf{x}}\|^2\right] \leq G^2$$

We also assume that we can efficiently project a point onto $\mathcal{K}$. i.e. given a point $\mathbf{y}$ we can compute

$$\Pi_{\mathcal{K}}(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x} - \mathbf{y}\|$$

The following fact is left as an exercise:

Fact 13.1. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set, $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} = \Pi_{\mathcal{K}}(\mathbf{y})$. Then for any $\mathbf{z} \in \mathcal{K}$ we have that:*

$$\|\mathbf{y} - \mathbf{z}\| \geq \|\mathbf{x} - \mathbf{z}\|$$

13.2.2 The SGD Algorithm

Algorithm 1 Stochastic Gradient Descent

0: **Input:** A convex function f, a convex domain $\mathcal{K}$ and a sequence of learning rates $\{\eta_t\}$.
for $t = 1, 2 \dots T$ **do**
Let $\hat{\nabla}_t = \mathcal{O}(\mathbf{x})$ and set $f_t(\mathbf{x}) = \langle \hat{\nabla}_t, \mathbf{x} \rangle$.
Update and project:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \hat{\nabla}_t$$
$$\mathbf{x}_t = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})$$

end for
return $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$.

Theorem 13.3. *Run Algorithm 1 with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ then*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_T)\right] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[f(\mathbf{x}_t)\right] \leq \min_{\mathbf{x}^* \in \mathcal{K}} f(\mathbf{x}^*) + \frac{3GD}{2\sqrt{T}}$$

13.2.3 Application to Learning Linear Classes

Before we dwell into the proof of Thm. 13.3, let us discuss its implication to learning convex problems. For concretness, let us focus on learning ℓ_2-norm constraint linear classifier:

Given a distribution over labeled examples and points from the unit ball $(\mathbf{x}, y) \sim D$, and a convex Lipschitz function, recall that we wish to solve the problem

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}_D(\mathbf{w}) = \mathbb{E}\left(\ell(\mathbf{w} \cdot \mathbf{x}, y)\right) \\ \text{s.t.} \quad & \|\mathbf{w}\| \leq B \end{aligned}$$

By linearity of the defivative we have that

$$\nabla \mathcal{L}_D(\mathbf{w}) = \nabla \mathbb{E}\left[\ell(\mathbf{w} \cdot \mathbf{x}, y)\right] = \mathbb{E}\left[\nabla \ell(\mathbf{w} \cdot \mathbf{x}, y)\right] = \mathbb{E}\left[\ell'(\mathbf{w} \cdot \mathbf{x}, y)\mathbf{x}\right]$$

In other word, by sampling IID examples $(\mathbf{x}, y)$ from D we can implement an oracle for a stochastic gradient thorugh

$$\mathcal{O}(\mathbf{x}) = \ell'(\mathbf{w} \cdot \mathbf{x}, y)\mathbf{x}$$

We can thus obtain the following application of SGD to learn linear classifiers:

Corollary 13.4. *Consider the convex problem $(\mathcal{X}, \mathcal{F}, \ell)$ where ℓ is L-Lipschitz convex loss $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$ and $\mathcal{F} = \{\mathbf{w} : \mathbf{w} : \mathbf{x} \to \mathbf{w} \cdot \mathbf{x},\ \|\mathbf{w}\| \leq B\}$. Given a access to $(\mathbf{x}, y) \sim D$ pairs distributed according to D, applying the stochastic graident oracle $\mathcal{O}(\mathbf{w}) = \hat{\nabla}_{\mathbf{w}} = \ell'(\mathbf{w} \cdot \mathbf{x}, y)\mathbf{x}$ with Alg. 1. After $T = O\left(\left(\frac{LB}{\epsilon}\right)^2\right)$ iterations we obtain in expectation:*

$$\mathbb{E}\left[\mathcal{L}_D\left(\bar{\mathbf{w}}_T\right)\right] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathbf{w}_t \cdot \mathbf{x}_t, y)\right] \leq \min_{\mathbf{w}^* \in \mathcal{K}} \mathcal{L}_D(\mathbf{w}^*) + \epsilon$$

Proof. Observing our estimator for the gradient $\hat{\nabla}_{\mathbf{w}}$, we can bound its second moment the Lipschitz constant, namely if $|\ell'(\mathbf{w}\cdot\mathbf{x}, y)| \leq L$ we have that for $\hat{\nabla}_{\mathbf{w}} = \ell'(\mathbf{w}\cdot\mathbf{x}, y)\mathbf{x}$:

$$\|\hat{\nabla}_{\mathbf{w}}\|^2 \leq L^2.$$

Thus: Given a sequence IID of labeled pairs $\{\mathbf{x}_i, y\}_{i=1}^T \sim D$, applying Alg. 1 to the sequence sequentially: We obtain an output $\bar{\mathbf{w}}_T$ such that by Thm. 13.3:

$$\mathbb{E}\left[\mathcal{L}_D\left(\bar{\mathbf{w}}_T\right)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_D(\mathbf{w}_t) \leq \min_{\mathbf{w}^*\in\mathcal{K}}\mathcal{L}_D(\mathbf{w}^*) + \frac{3LB}{2\sqrt{T}}$$

If $T \geq \frac{4(LB)^2}{9\epsilon^2}$ we obtain:

$$\mathbb{E}\left[\mathcal{L}_D\left(\bar{\mathbf{w}}_T\right)\right] \leq \min_{\mathbf{w}^*\in\mathcal{K}}\mathcal{L}_D(\mathbf{w}^*) + \epsilon$$

□

13.2.4 High Probability Rates

For PAC learning, we wish need to show that if we apply SGD w.p $1-\delta$ we obtain

$$\mathcal{L}_D\left(\bar{\mathbf{w}}_T\right) \leq \min_{\mathbf{w}^*\in\mathcal{K}}\mathcal{L}_D(\mathbf{w}^*) + \epsilon$$

We will assume for simplicity that $|\ell(0, y)| \leq LB$. We again use a concentration bound to pass from a bound in expectation to a bound w.h.p. Here we use Azuma's inequality for martinagles. Namely

Azuma's inequality Let $Z_1, \ldots, Z_m$ be a sequence of martinagles i.e. $\mathbb{E}\left[Z_i | Z_1, \ldots, Z_{i-1}\right] = 0$, bounded by B, then for all $\epsilon > 0$ we have that

$$\mathbb{P}(\sum Z_i \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2B^2 m}}$$

Note that the sequence $\ell(\mathbf{w}_t\cdot\mathbf{x}_t, y) - \mathcal{L}_D(\mathbf{w}_t)$ is indeed a martinagle (as $\mathbf{w}_t$ is deterministic given the past and $(\mathbf{x}_t, y_t)$ is independent of the past). Further, $|\mathbf{w}\cdot\mathbf{x}| \leq B$ hence by Lipschitzness $\ell(\mathbf{x}_t, y_t) \leq LB + \ell(\mathbf{w}\cdot\mathbf{x}, y) \leq 2LB$. We can thus obtain that

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\ell(\mathbf{w}_t\cdot\mathbf{x}_t, y) - \mathcal{L}_D(\mathbf{w}_t) \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{8(LB)^2 m}}$$

Thus we obtain that if $m > O((LB)^2\frac{\log 1/\delta}{\epsilon^2})$ then with probability $1-\delta$ over the sequence $\mathbf{w}_1, \ldots, \mathbf{w}_T$:

$$\mathcal{L}_D(\bar{\mathbf{w}}_T) \leq \frac{1}{T}\sum_{t=1}^{T}\ell(\mathbf{w}_t\cdot\mathbf{x}_t, y_t) \leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_D(\mathbf{w}_t) + \epsilon \leq 2\epsilon$$

13.2.5 Do we need Rademacher Complexity???

A striking fact is that we obtained generalization bounds for SGD without having to invoke a Rademacher bound!!! SGD indeed comes with its own generalization guarantee which is comparable (up to constant factors) to the bound obtained using Rademahcer complexity.

There are two main differences between the two generalization bound we obtain: Rademacher Complexity guarantee that no matter what algorithm we use and no matter which target function we get: The empirical loss is close to the expected loss – i.e. uniform convergence!!! The bound from SGD, in contrast, does not guarantee uniform convergence and is applicable only for SGD algorithm: Meaning if we use the SGD algorithm we will learn.

Both bounds have different applications. In practice, sometimes, one uses SGD with repetition over the sample point (i.e. we generate a sequence of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ and then draw random examples from the empirical data. Rademacher bound will apply for this setting also, while the bound for SGD holds w.r.t to the empirical distribution and is inapplicable for the expected loss.

On the other hand, the SGD bound is true even in some cases where there is no uniform convergence hence Rademahcer bounds are inapplicable. In [1], the authros showed a learning problem where, via stochastic optimization one can learn the problem, yet there is no uniform convergence.

References

[1] Shalev-Shwartz, Shai, and Sahmir Ohad, and Srebro Nathan, and Sridharan Karthik. *Stochastic Convex Optimization.* Conference on Learning Theory (2009)