The Algorithmic Foundations of Adaptive Data Analysis September 12, 2017

Lecture 2: Uniform Convergence and Optimization

Lecturer: Adam Smith *Scribe: Adam Smith*

1 A Model for Adaptive Data Analysis

This course will focus on settings where data analysis is interactive, and questions asked by the analysis be depend *adaptively* on answers to previous questions. A stylized but important setting for thinking about adaptivity is that of *statistical queries*. Recall that a statistical query asks for the expectation of a bounded function in the population. Such queries capture a wide range of basic descriptive statistics (the prevalence of a disease in a population, for example, or the average age). Many inference algorithms can also be expressed in terms of a sequence of statistical queries [Kea98].

Suppose each data point lies in a universe $\mathcal{X}$, so that a data set lies in $\mathcal{X}^n$ and the underlying population is a distribution on $\mathcal{X}$. A *statistical query* is specified by a function $\phi : \mathcal{X} \to [0,1]$. The *population value* (or *population mean* or *true value*) of a linear query is its expected value when evaluated on a fresh sample from $\mathcal{D}$, denoted (with some abuse of notation)

$$\phi(\mathcal{D}) \stackrel{\text{def}}{=} \mathbb{E}_{X\sim\mathcal{D}}[\phi(X)]\,.$$

Consider now an analyst A who whishes to answer a sequence such queries $\phi_1, \phi_2, ...$ For each query ϕ_j, the analyst wishes to learn the population mean $\phi_j(\mathcal{D})$ as accurately as possible.

The analyst A typically does not have direct access to $\mathcal{D}$, however, so instead makes use of a data set $\mathbf{S} = (X_1, ..., X_n)$ of n points drawn i.i.d. from $\mathcal{D}$. The most straightforward way to estimate $\phi_j(\mathcal{D})$ is via the *empirical mean* on $\mathbf{S}$:

$$\phi(\mathbf{S}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{x_i\in\mathbf{s}} \phi(X_i)\,.$$

Next lecture, we will see settigns where using the empirical mean to estimate the population mean is *not* the best strategy! To allow ourselves more flexibility, we will imagine that the analyst interacts with the data via a "mechanism" M. Formally, given the query answering mechanism M, a data analyst A, and a distribution $\mathcal{D}$ on the data universe $\mathcal{X}$, we consider a random interaction defined by selecting a sample $\mathbf{S}$ of n i.i.d. draws from $\mathcal{D}$, and then having A interact with $M(\mathbf{S})$ for k rounds, where in each round j, (i) A selects ϕ_j (based on $a_1, \ldots, a_{i-1}$), (ii) M answers a_j.

This interaction is illustrated in Figure 1.

In general, neither the mechanism nor the analyst knows the exact distribution $\mathcal{D}$ (otherwise, why collect data?), so the mechanism cannot always answer with the population mean $\phi(\mathcal{D})$. The naïve mechanism that always returns the empirical mean (that is, for which $a_j = \phi_j(\mathbf{S})$) is called the *empirical mechanism*. Here are a few other examples of mechanisms we might use (the list is by no means exhaustive!):

1. Rounding: report $\phi_j(\mathbf{S})$ rounded to the nearest multiple of 0.1 (0r 0.01 or...);
2. Noise addition: report $\phi_j(\mathbf{S}) + Z_j$ where $Z_j \sim N(0, \sigma^2)$ (for fixed $\sigma > 0$);
3. Subsampling: for each ϕ_j, take a subsample $\mathbf{S}_j \subseteq \mathbf{S}$ and report $\phi_j(\mathbf{S})$;
4. Given a Bayesian prior on the distribution $\mathcal{D}$, construct a posterior distribution on $\mathcal{D}$ given $\mathbf{S}$, and answer each query ϕ_j using the expectation over the posterior.

How do we measure the mechanism's performance? For now, we will the meachism's worst absolute error, as measured with respect to the population: The *(population) error* of M is the random variable

$$\text{err}_{\mathbf{S}}(M, A) = \max_j |\phi_j(\mathcal{D}) - a_j|\,.$$

which depends on $\mathbf{S}$ as well as the coins of M and A.

Definition 1 *A query answering mechanism $\mathcal{M}$ is (α, β)-accurate on i.i.d. data for k queries if for every data analyst A and distribution $\mathcal{P}$, we have*

$$\Pr\left(\mathrm{err}_{\mathbf{S}}(M, A) \leq \alpha\right) \geq 1 - \beta\,.$$

The probability is over the choice of the dataset $\mathbf{S} \sim_{i.i.d.} \mathcal{D}$ and the randomness of the mechanism and the analyst. Similarly, the expected error *of M is the supremum, over distributions $\mathcal{D}$ and data analysts A, of*

$$\mathbb{E}(\mathrm{err}_{\mathbf{S}}(M, A))\,.$$

We sometimes fix the distribution $\mathcal{D}$ and take the supremum only over analysts A.

The definition makes no assumptions on how the analyst selects queries, except that the selection is based on the outputs of M and not directly on the data.

Phrased somewhat differently, we are interested in an unusual minimax problem, where the "max" includes all possible analyst strategies. For the case of expected error, this corresponds to

$$\inf_{\substack{\text{mechanisms}\\ M}} \left(\sup_{\substack{\text{distributions}\\ \mathcal{D}}} \sup_{\substack{\text{analysts}\\ A}} \mathop{\mathbb{E}}_{\substack{\mathbf{S}\sim\mathcal{D}\\ i.i.d.}} \left(\mathrm{err}_{\mathbf{S}}(M, A)\right) \right).$$

Before trying to understand this general setting, though, let's spend a bit more time on the simplest strategy, in which all queries are specified ahead of time.

2 Answering Nonadaptive Statistical Queries

Last time, we discussed what happens when we try to estimate the expectation of a single statistical query[1] using a sample of size n drawn i.i.d from a distribution $\mathcal{D}$.

Theorem 2 *Let $\mathcal{D}$ be a distribution on the set $\mathcal{X}$, and $\phi : \mathcal{X} \to [0,1]$ be a statistical query with expectation $\mu = \mathbb{E}_{\mathcal{D}}[\phi]$. If $S \sim \mathcal{D}^n$ is a sample of size n drawn i.i.d. from $\mathcal{D}$, then, probability $1-\delta$ over the choice of S: $|\mathbb{E}_{\mathbf{S}}[\phi] - \mu| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}$.*

If we want the empirical mean to be within α of the true expectation with probability $1-\delta$, a sample of size $\frac{\ln(1/\delta)}{2\alpha^2}$ thus suffices—or $O\left(\frac{1}{\alpha^2}\right)$ when δ is constant.

[1] Terminology: What computer scientists call statistical queries, statisticians call *bounded linear functionals*. A *functional* in this context is a map from probability distributions to real numbers (in this case, ϕ maps $\mathcal{D}$ to $\mathbb{E}_{\mathcal{D}}[\phi]$). *Linear* means that the functional's value is its expectation on a single data point selected according to the distribution, and *bounded* refers to the statistic taking values in $[0,1]$ (or any other bounded interval).

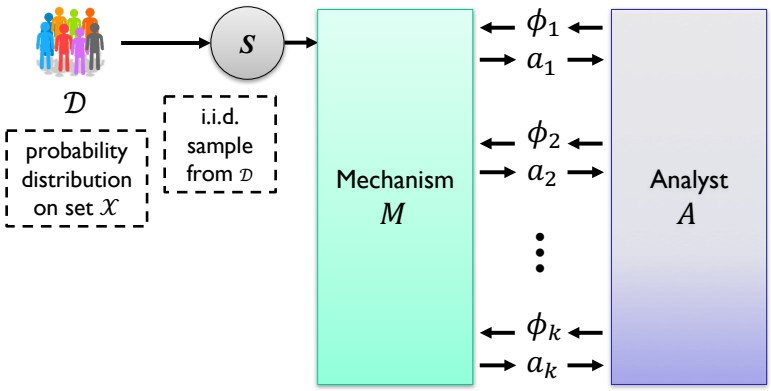

Figure 1: Adaptively selected linear queries

Most interesting analyses ask more than one question of the data. In the simplest setting, the analyst has a set of questions in mind that were specified *before the data were collected*—what we will call "nonadaptively".

Suppose that an analyst specifies k statistical queries $\phi_1, ..., \phi_k$ ahead of time. How many samples does it take to estimate the expectations of all of these simulatenously?

Theorem 3 (Nonadaptive statistical queries) *Let $\mathcal{D}$ be a distribution on the set $\mathcal{X}$, and $\phi_1, ..., \phi_k : \mathcal{X} \to [0,1]$ be statistical queries with expectations $\mu_j = \mathbb{E}_{\mathcal{D}}[\phi_j]$. If $S \sim \mathcal{D}^n$ is a sample of size n drawn i.i.d. from $\mathcal{D}$, then, probability $1-\delta$ over the choice of S: $\max_{j=1,...,k} |\mathbb{E}_{\mathbf{S}}[\phi_j] - \mu_j| \leq \sqrt{\frac{\ln(2k/\delta)}{2n}}$.*

Proof Recall the union bound: for every set of k events $E_1, ..., E_k$ in the same probability space, the probability of their union is at most the sum of their probabilities:

$$\Pr\left(\bigcup_{j=1,...,k} E_j\right) \leq \sum_{j=1}^{k} \Pr(E_j)\,.$$

By Theorem 2, it is unlikely that the expectation of any particular query will be way off:

$$\forall i \in \{1, ..., k\}: \quad \Pr\left(|\mathbb{E}_{\mathbf{S}}[\phi_j] - \mu_j| > \sqrt{\frac{\ln(2k/\delta)}{2n}}\right) \leq \frac{\delta}{k}\,.$$

So the probability that any one of the empirical means of the k of the queries will be far from their true expectations is at most δ:

$$\Pr\left(\forall i \in \{1, ..., k\}: \quad |\mathbb{E}_{\mathbf{S}}[\phi_j] - \mu_j| > \sqrt{\frac{\ln(2k/\delta)}{2n}}\right) \leq \delta\,.$$

Considering the complementary event yields the theorem statement. ■

Exercise 1 *If we want the empirical mean of all k queries to be within α of the true expectation with probability $2/3$, what sample size suffices asymptotically?*

Exercise 2 *Prove that $\mathbb{E}\left(\max_{j=1,...,k} |\phi_j(\mathbf{S}) - \mu_j|\right) = O(\sqrt{\frac{\ln(2k)}{2n}})$.*

Suppose that we are only interested in finding out the *smallest* of the expectations of the ϕ_j. That is, we wish to estimate $\mu_{min} = \min_{j=1,...,k} mu_j$. Alternatively, we may wish to obtain an approximate minimizer—that is, to find an index $\hat{\imath}$ such that $\mathbb{E}_{\mathbf{S}}[\phi_{\hat{\imath}}] - \mu_{min}$ is as small as possible. In such settings, the bound of the Theorem 3 applies up to an additional factor of 2 (why?): With probability at least $1-\delta$, we have

$$|\mathbb{E}_{\mathbf{S}}[\phi_{\hat{\imath}}] - \mu_{min}| \leq 2\sqrt{\frac{\ln(2k/\delta)}{2n}} \text{ where } \hat{\imath} = \arg\min_j \mathbb{E}_{\mathbf{S}}[\phi_j]\,. \tag{1}$$

Exercise 3 *Show that the bound of Equation (1) is tight in general. That is, there exists a distribution and a collection of statistical queries such that the probabilities in (1) are $\Omega\left(\sqrt{\frac{\ln(k/\delta)}{n}}\right)$ for large n, k and small δ.*

3 Uniform Convergence and Optimization

Theorem 3 is probably the simplest example of *uniform convergence* of random variables, where we ask for the probability that a whole set of random variables are simultaneously close to their expectations.

There are thousands of beautiful results on concentration, many of which rely on something like Theorem 3 at their core. We will see one or two more examples this lecture.

3.1 Stochastic Optimization

A lot of statistical problems involve finding "likely" parameters of some probability model, given a set of observed data. For example, in the classic "ordinary least squares" linear regression problem, every data point is a pair (x, y) in $\mathbb{R}^d \times \mathbb{R}$, where the entries of x are called the features or independent variables, and y is the response. A common task is to look for a linear relationship between x and y by looking for a vector $w \in \mathbb{R}^d$ that minimizes the mean squared error

$$\ell_{OLS}(w; S) = \frac{1}{n} \sum_{(x_i, y_i) \in S} \left(y_i - w^T x_i\right)^2.$$

This corresponds to computing the maximum likelihood estimator for a model in which there is a hidden vector w^* such that each y_i is generated as $w^T x_i + Z_i$ and the Z_i' are i.i.d. $N(0, \sigma^2)$ for some fixed $\sigma > 0$.

If we consider the expected value of $\ell_{OLS}(w; S)$ over the choice of S, we get the *population error* $L(w; \mathcal{D}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left(Y - w^T X\right)^2$. This achieves its minimum value, σ^2, at w^* if $\mathcal{D}$ really follows the model.

More generally, we imagine a parameter space w—typically a subset of $\mathbb{R}^d$ for finite d—in which every parameter vector w and data point x are associated with a loss $\ell(w; x)$ describing how "well" w matches x. Given a data set $S = (x_1, ..., x_n)$ drawn i.i.d. from a distribution $\mathcal{D}$, we have two key quantities:

$$\text{Emprical loss:} \qquad \ell(w; S) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(w; x_i)\,, \tag{2}$$

$$\text{Population loss:} \qquad L(w; \mathcal{D}) \stackrel{\text{def}}{=} \mathbb{E}_{S\sim\mathcal{D}^{\otimes n}}[\ell(w; S)] = \mathbb{E}_{x\sim\mathcal{D}}[\ell(w; x)]\,. \tag{3}$$

Our goal is generally to approximate the *population minimizer*

$$w^* \stackrel{\text{def}}{=} \arg\min_{w \in w} L(w; \mathcal{D})\,.$$

We'll measure how well we do with a given estimate $\hat{w}$ by the excess risk

$$\text{err}_{\mathcal{D}}(\hat{w}) = L(\hat{w}; \mathcal{D}) - \min_{w \in w} L(w; \mathcal{D})\,.$$

A commonly used estimator is the empirical minimizer

$$\hat{w}_{erm} \stackrel{\text{def}}{=} \arg\min \ell(w; S)\,.$$

This setup captures a class of statistical estimators called *M-estimators*. For least-squares regression, w denotes the regression coefficents and ℓ denotes the mean squared error; for classification based on a deep neural network, w might be the weights in the neural network and ℓ, the probability of correct classification

Exercise 4 *Suppose $\mathcal{X} = w = \mathbb{R}$ and $\ell(w; x) = |w - x|$. What well-known function of S is given by the empirical minimizer $\hat{w}_{emd}$? In terms of the distribution $\mathcal{D}$, what is w^*?*

How well does the empirical minimizer do at minimizing the population error? In general, in this setup, it can be arbitrarily horrible. However, under mild assumptions, it can be shown to perform very well! In some settings it can also be modified to perform even better via "regularization"—we'll get to that later in the class.

Assumption 4 (Lipschitz loss on a bounded parameter space) *We assume that: (a) $\Theta \in \mathbb{R}^d$ is contained in a ball of radius R (that is, $\|w\| \le R$ for all $w \in \Theta$, where $\|\cdot\|$ denotes the usual Euclidean norm). (b) the loss doesn't jump around too quickly as w changes. More specifically, ℓ is C-Lipschitz: for all $u, v \in \Theta$ and all $x \in \mathcal{X}$, we have $|L(u; x) - L(v; x)| \le C\|u - v\|$. When $\ell(\cdot; x)$ has a gradient, this is the same as assuming that its gradient has norm at most C everywhere in θ.*

Theorem 5 (Uniform convergence) *Under Assumption 4, if S is drawn i.i.d. from $\mathcal{D}$, then for all suficiently large n, with probability at least $1-\delta$ over S, $\sup_{w\in\Theta}|\ell(w;S)-L(w;\mathcal{D})| \le 6RC\sqrt{\frac{d\log(n/\delta)}{n}}$.*

Corollary 6 *Under Assumption 4, if S is drawn i.i.d. from $\mathcal{D}$, then $\mathrm{err}(w_{erm}) \le 6RC\sqrt{\frac{d\log(n/\delta)}{n}}$ with probability at least $1-\delta$ over S.*

We will prove this using a "net argument".

Definition 7 (Cover) *Given a set $\Theta \subseteq \mathbb{R}^d$, a subset $N \subseteq \Theta$ is an α-cover (or α-net) of Θ if, for every $u \in \Theta$, there is a point $v \in N$ within distance at most α of u.*

Lemma 8 *For every set $\Theta \in \mathbb{R}^d$ contained in a ball of radius R, for every $\alpha > 0$, there is an α-cover of size at most*

$$\left(1+\frac{2R}{\alpha}\right)^d.$$

Proof [of Lemma 8] First, note that we can take $R=1$ without loss of generality, since an $\frac{\alpha}{R}$-cover of $\frac{1}{R}\cdot\Theta$ can be scaled up to an α-cover of Θ.

We'll proceed via a standard volume argument. Fix $\alpha > 0$ and choose N_α to be a maximal α-separated subset of Θ. In other words, N_α is such that $\|u-v\| \ge \alpha$ for all $u, v, \in N_\alpha$, $u \ne v$, and no subset of Θ containing N_α has this property.

The maximality property implies that N_α is an α-cover of Θ. Indeed, otherwise there would exist $u \in \Theta$ that is at least α-far from all points in N_α. So $N_\alpha\{x\}$ would still be an α-separated set, contradicting the minimality property.

This is where the volume argument comes in. First, the separation property implies that the balls of radii $\alpha/2$ centered at the points in N_α are disjoint, so the volume of their union is the some of their volumes. On the other hand, all such balls lie in $(1+\alpha/2)B$ where B denotes the unit Euclidean ball centered at the origin. Comparing volumes gives

$$\underbrace{vol\Big(\frac{\alpha}{2}B\Big)\cdot|N_\alpha|}_{\text{vol. of union of } \alpha/2\text{-balls around } N_\alpha} \le vol\Big((1+\frac{\alpha}{2})B\Big).$$

Since $vol(rB) = r^d vol(B)$ for all $r \ge 0$, we conclude that $|N_\alpha| \le (1+\frac{\alpha}{2})^d/(\frac{\alpha}{2})^d = (1+\frac{2}{\alpha})^d$ as required. ■

Proof [of Theorem 5] Consider a net $N_\alpha = \{w_1, ..., w_K\}$ of Θ as in the previous lemma, and let $K \le (1+2R/\alpha)^d$ be its size. We will set α later.

We will assume for simplicity that, for each $x \in \mathcal{X}$, the minimum value of $\ell(\cdot;x)$ on Θ is 0 (we can enforce this condition by adding or subtracting a function of x to ℓ; this will not affect minimization). In particular, this means that ℓ will take only values in $[0, 2CR]$ since ℓ is C-Lipschitz in w, and Θ has diameter at most $2R$.

For each point $w_j \in N_\alpha$, we can think of $\ell(w;\cdot)$ as decribing a statistical query (that takes values in $[0,2CR]$ instead of $[0,1]$) with expected value $L(w;\mathcal{D})$. Applying an appropriately scaled version of Theorem 3, we get that with probability $1-\delta$,

$$\max_{w\in N_\alpha}|\ell(w;S)-L(w;\mathcal{D})| \le 4CR\sqrt{\frac{\ln(2K/\delta)}{2n}}.$$

This statement takes care of the cover. To extend the statement to every point in Θ, first observe that the average of Lipschitz functions is itself Lipschitz; thus, $\ell(\cdot;S)$ and $L(\cdot,\mathcal{D})$ are Lipschitz. Now for any $w \in \Theta$, we can find a cover point w_j within distance at most α of w. By the triangle inequality,

$$|\ell(w;S)-L(w;\mathcal{D})| \le |\ell(w;S)-\ell(\tilde{w};S)| + |\ell(\tilde{w};S)-L(\tilde{w};\mathcal{D})| + |L(\tilde{w};\mathcal{D})-L(w;\mathcal{D})| \quad (4)$$

$$\le C\alpha + 2CR\sqrt{\frac{\ln(2K/\delta)}{2n}} + C\alpha \quad (5)$$

$$\le 2C\alpha + 4CR\sqrt{\frac{d\ln(1+2R/\alpha)+\ln(2/\delta)}{2n}} \quad (6)$$

where the last inequality comes from the bound on K. Setting $\alpha = R/n$ yields the desired bound for sufficiently large n (recalling that CR/n is always smaller than $CR/\sqrt{n}$, and that $4/\sqrt{2} < 3$). ■

Exercise 5 *Prove Corollary 6.*

Where does this leave us? If we can find the empirical minimum w_{erm} efficiently, then—assumging $n \gg d$—we are guaranteed to get a value which is approximately optimal for the distribution. How difficult the empirical optimization problem is varies enormously from context to context, but one large and important class of problems have loss functions that are *convex*. The structure of convex functions make them amenable to greedy optimization methods, such as gradient descent.

3.2 Net arguments in general

The basic outline of the previous proof is very common in probability theory: to prove that a family of ranodm variables covnerges uniformly to their expectation, we first find a subset that is somehow "representative"—inasmuch as every random variable is close to one of the representative ones. We then show by a union bound that the representative random variables converge simultaneously at some rate, and conclude convergence for the entire family.

There are lots of settings where such simple strategies are not enough, however, and many concepts have been developed to understand uniform convergence; Vapnik-Chernovenkis (VC) dimension is a famous example. We may return to these later in the class.

4 Convex Optimization

What good are statistical queries? As mentioned above, many learning algorithms can be expressed as a sequence of statistical queries. A fine example of this was just mentioned above—the problem of finding a minimizer for a convex function.

Recall that a set $\Theta \subseteq \mathbb{R}^d$ is *convex* if for every $x, y \in \Theta$, the line segment $x\bar{y}$ is contained in Θ. A function $\Theta \to \mathbb{R}$ is convex if it is "curved upwards" everywhere. There are many equivalent definitions of this. The simplest requires that, for all $x, y \in \Theta$,

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}\,.$$

It is often easier to work with the following equivalent definition: a function f is convex of Θ if for every $x \in \Theta$, there exists a vector $g \in \mathbb{R}^d$, called a *subgradient for f at x*, such that the linear function $\hat{f}(y) = f(x) + \langle g, y - x\rangle$ is a lower bound for f on Θ (that is, $f(y) \geq \hat{f}(y)$ for all $y \in \Theta$). There can be many subgradients at a given point; the set of all subgradients is denoted $\partial f(x)$. When the subgradient is unique, it coincides with the usual gradient $\nabla f(x)$.

Convex functions have the feature that they have no local minima (Why?). As a consquence, simple greedy strategies will actually find global optima. For example, we may consider projected gradient descent, in which we start with an arbitrary value $x_0 \in \Theta$, and derive estimates x_t for $t = 1, 2, ...$ using

$$\begin{aligned} y_{t+1} &= x_t - \eta g_t, \text{ where } g_t \in \partial f(x_t) \\ x_{t+1} &= \Pi_\Theta(y_t) \end{aligned}$$

where the Π_Θ operator is the *projection* onto Π, i.e. $\Pi_\Theta(y) = \arg\min_{x\in\Theta} \|x - y\|$.

Lemma 9 *Let Θ be closed and convex. Then*

1. *For every $y \in \mathbb{R}^d$, the projection $\Pi_\Theta(y)$ is unique.*
2. *For every $x \in \Theta$ and $y \in \mathbb{R}^d$,*
$$\langle \Pi_\Theta(y) - x,\ \Pi_\Theta(y) - y\rangle \leq 0\,.$$
In particular, $\|\Pi_\Theta(y) - x\| \leq \|y - x\|$ (that is, projection decreases the distance to all points in Θ).

Theorem 10 *Let Θ be contained in a ball of radius R and f be a C-Lipschitz convex function on Θ. If we run T rounds of projected gradient descent with $\eta = \frac{R}{C\sqrt{T}}$, then*

$$f\left(\frac{1}{T}\sum_{j=1}^{T} x_t\right) - f(x^*) \le RC/\sqrt{T}\,, \qquad \textit{where} \qquad x^* \stackrel{def}{=} \arg\min_{x\in\Theta} f(x)\,.$$

Proof We will consider the distance $f(x_t) - f(x^*)$. Since f is convex, we can bound this distance using the linear approximation to f at x_t:

$$f(x_t) - f(x^*) \le \langle g_t, x_t - x^*\rangle\,.$$

By construction, the subgradient g_t is exactly the difference $\frac{1}{\eta}(y_{t+1} - x_t)$. Using the identity $2\langle u, v\rangle = \|u\|^2 - \|v\|^2 - \|u - v\|^2$, we get:

$$\begin{aligned} f(x_t) - f(x^*) &\le \frac{1}{\eta}\langle y_{t+1} - x_t, x_t - x^*\rangle \\ &= \frac{1}{\eta}\big(\|y_{t+1} - x_t\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\big) \\ &= \frac{1}{\eta}\big(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\big) + \eta\|g_t\|^2\,. \end{aligned}$$

Recall that projection only decreases distances, so $\|y_{t+1} - x^*\|^2 \ge \|x_{t+1} - x^*\|^2$ and

$$f(x_t) - f(x^*) \le \frac{1}{\eta}\big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\big) + \eta C^2\,. \tag{7}$$

Finally, we see that the quantity we are trying to bound for the theorem is the average of the left-hand side of (7) (over t from 0 to $T-1$). Moreover, the right-hand side telescopes!

$$f\left(\frac{1}{T}\sum_{j=1}^{T} x_t\right) - f(x^*) \le \frac{1}{\eta}\big(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\big) + \eta T C^2 \le \frac{R^2}{T\eta} + \eta C^2\,.$$

Setting $\eta = \frac{R}{C\sqrt{T}}$ proves the theorem. ■

4.1 Optimizing the Population Loss

We can use Theorem 10 together with our uniform convergence statement (Theorem 5) to get a nice statistical query algorithm: use gradient descent to optimizie the empirical mean, and then conclude that the resulting optimum is pretty close to perfect via uniform convergence.

This is a statistical query algorithm because each coordinate of each of the queries to the empirical loss's gradient is in fact a statistical query (with values in $[-C, C]$ instead of $[0, 1]$). Moreover, this algorithm is really adaptive: the query point of gradient descent depends on the gradients of previous rounds. However, our analysis takes advantage of the fact that there are only so many places the algorithm can wander.

As the course progresses, we'll see (a) a better analysis of this type of gradient descent that allows us to get much tighter bounds on the excess loss, and (b) that not all adaptive strategies for asking statistical queries are so benign.

Notes

The proof of Lemma 8 comes from `http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf`

Our analysis of gradient descent follows Section 3.1 of `https://arxiv.org/pdf/1405.4980.pdf`, which follows the book of Nesterov.

References

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.