

Lecture Differentially Private Empirical Risk Minimization

Aleksandar Nikolov

Scribe: Patricia Thaine

1 Set up

Today we will talk about differentially private machine learning, and specifically, about supervised learning in an agnostic context, when no concept class may fit all the labels. This topic will also give us a chance to discuss efficient differentially private optimization algorithms, in particular, private gradient descent.

Our dataset X will consist of n labeled points, i.e., $X = \{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$. Here \mathcal{X} is the set of possible data points, and \mathcal{Y} is the set of possible labels. We will focus on d -dimensional data, i.e. $\mathcal{X} = [0, 1]^d$ or $\{0, 1\}^d$, and binary labels, i.e. $\mathcal{Y} = \{\pm 1\}$. We usually think of each (x_i, y_i) as coming from some unknown distribution D over \mathcal{X} .

Given a concept class $C = \{c_\theta : \theta \in \Theta\}$, our goal will be to find the function c_θ (or, equivalently, the parameter $\theta \in \Theta$) that best fits the data distribution D . This will be modeled as minimizing a loss function $\ell : \Theta \times (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$. A natural example is the binary loss:

$$\ell(\theta, (x, y)) = \begin{cases} 1 & c_\theta(x) \neq y \\ 0 & c_\theta(x) = y \end{cases}$$

Then if $c_\theta(x) = y$, i.e. c_θ correctly predicts the label of x , the loss $\ell(\theta, (x_i, y_i))$ is 0, and otherwise it is 1. Our goal is to minimize the loss $L_D(\theta) = \mathbb{E}_{(x,y) \sim D}[\ell(\theta, (x, y))]$. Usually, we instead minimize the empirical loss (or empirical risk) as a proxy, i.e., we minimize

$$L_X(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, (x_i, y_i)).$$

We hope that minimizing $L_X(\theta)$ corresponds to approximately minimizing $L_D(\theta)$. Statistical learning theory studies conditions on the concept class C , the loss ℓ , and the distribution D , under which we can guarantee this convergence. While there are many interesting questions there, they are beyond the scope of this lecture. Instead, we will focus on how to minimize $L_X(\theta)$ efficiently under differential privacy.

We will focus on a classical example of a family of functions: the linear predictors

$$c_\theta(x) = \text{sign}\left(\sum x_j \theta_j + \theta_0\right), \theta \in \mathbb{R}^{d+1}$$

Then $\sum x_j \theta_j + \theta_0 = 0$ is the equation of a hyperplane, and $f_\theta(x) = +1$ “above” the hyperplane, and $f_\theta(x) = -1$ below it. We have the following picture when $d = 2$:

$$\begin{array}{l} +1 \\ \bullet(x, y) = \begin{cases} y = +1, & \text{loss is } 0 \\ y = -1, & \text{loss is } 1 \end{cases} \\ -1 \end{array}$$

2 (Constrained) Logistic regression

While the binary loss from our example so far is very natural, it makes minimizing $L_X(\theta)$ computationally hard. A way around this is to define a surrogate loss for which ERM is more tractable. We will focus on the *logistic loss* function, defined as

$$\ell(\theta, (x, y)) = \log(1 + e^{-c_\theta(x) \cdot y}),$$

and the class of linear functions $c_\theta(x) = \sum_{i=1}^d \theta_i x_i + \theta_0$ over $x \in \mathcal{X} = [0, 1]^d$. By adding a coordinate $x_0 = 1$ to every point x (so expanding \mathcal{X} to $[0, 1]^{d+1}$), we can simplify c_θ a bit to just $c_\theta(x) = \langle \theta, x \rangle$. This set up is motivated by a statistical model in which the label y of a point x is a random variable with a distribution parametrized by θ , and the logarithm of the odds ratio $\frac{\mathbb{P}(y|\theta)}{\mathbb{P}(-y|\theta)}$ is $\langle \theta, x \rangle$. Then $\frac{1}{1+e^{-\langle \theta, x \rangle}}$ is interpreted as $\mathbb{P}(y|\theta)$, and minimizing the logistic loss corresponds to finding the maximum likelihood estimate of θ given the data.

We will solve a constrained version of this problem in which $\theta \in \Theta = B_2^{d+1}(R)$, where $B_2^{d+1}(R)$ is the Euclidean ball of radius R centered at 0. This is a form of regularization: it constrains the hypothesis c_θ we find to be “simple”, thus helping with generalization. For us, this constraint will be crucial so that we can solve the ERM problem with differential privacy.

Instead of imposing a constraint, we can instead work with the ℓ_2 -regularized logistic loss $\ell(\theta, (x, y)) = \log(1 + e^{-\langle \theta, x \rangle \cdot y}) + \frac{\lambda}{2} \cdot \|\theta\|_2^2$. This is similar to the constrained problem, as it forces the optimal solution to be in a ball of bounded radius. The techniques we will use adapt easily to the regularized setting as well.

While we will use logistic regression as a running example in this lecture, the methods apply much more generally to constrained empirical risk minimization with convex loss.

3 Noisy Gradient Descent

Unlike least squares regression, there is no simple closed form expression for the optimal θ in the logistic regression problem. Instead we usually solve it (approximately) using a general purpose convex optimization algorithm. A popular choice is gradient descent:

```
 $\theta_0 \leftarrow 0$   
for  $t = 1 \dots T - 1$  do  
   $\tilde{\theta}_t \leftarrow \theta_{t-1} - \eta \nabla L(\theta_{t-1}, X)$   
   $\theta_t \leftarrow \tilde{\theta}_t / \max\{1, \|\tilde{\theta}_t\|_2 / R\}$   
end for  
output  $\frac{1}{T} \sum_{t=0}^{T-1} \theta_t$ 
```

Above, $\nabla L_X(\theta)$ is the gradient with respect to θ , i.e. it’s a vector whose i -th coordinate is the partial derivative $(\nabla L_X(\theta))_i = \frac{\partial L(\theta, X)}{\partial \theta_i}$. At every step, the algorithm moves θ a little bit in the direction opposite to the gradient: this is the direction in which the loss locally decreases the fastest. If θ ever leaves the ball, it is scaled back inside. See Figure 1.

Our approach will be to try to make this algorithm differentially private. Notice that, in any single step, the only thing that depends on the database X is the gradient of the loss L . So, a natural strategy is to add noise to the gradient. We get the following noisy variant of gradient descent:

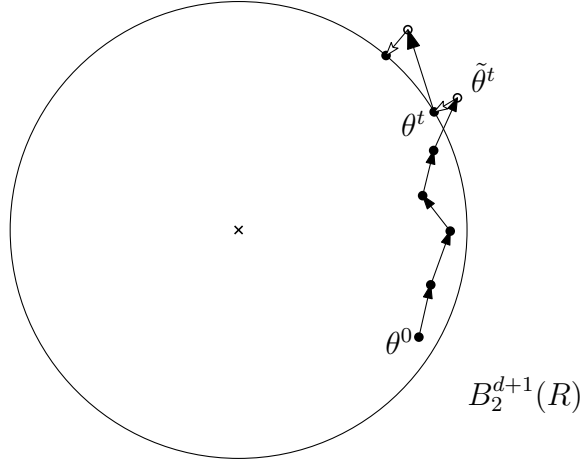


Figure 1: Gradient Descent

```

 $\theta_0 \leftarrow 0$ 
for  $t = 1 \dots T - 1$  do
   $\tilde{\theta}_t \leftarrow \theta_{t-1} - \eta(\nabla L(\theta_{t-1}, X) + Z_t)$ 
   $\theta_t \leftarrow \tilde{\theta}_t / \max\{1, \|\theta_t\|_2/R\}$ 
end for
output  $\frac{1}{T} \sum_{t=0}^{T-1} \theta_t$ 

```

Z_t is random noise that is added to make gradient descent differentially private. The noise will be Gaussian, and in the next section we will explore how large it needs to be.

4 Advanced composition (for Gaussian noise)

Recall the Gaussian noise mechanism: for a function f that maps databases to m -dimensional vectors, we release $\mathcal{M}_{\text{Gauss}}(x) = f(X) + Z$ where $Z \sim \mathcal{N}\left(0, \frac{(\Delta_2 f)^2}{\rho} \cdot I\right)$, i.e., every coordinate of Z is an independent Gaussian with mean 0 and $\frac{(\Delta_2 f)^2}{\rho}$. Here $\Delta_2 f$ is the ℓ_2 -sensitivity of f . For any $\delta > 0$, $\mathcal{M}_{\text{Gauss}}$ is (ε, δ) -DP for $\varepsilon = \rho + \sqrt{2\rho \ln(1/\delta)}$, so setting $\rho = \rho_{\varepsilon, \delta} \approx \frac{\varepsilon^2}{\ln(1/\delta)}$ gives an (ε, δ) -DP mechanism.

Say we want to publish k functions f_1, \dots, f_k such that $\Delta_2 f_i \leq C$ for each one of them. One thing we can do is publish each one using the Gaussian noise mechanism with privacy parameters $(\varepsilon/k, \delta/k)$, and then use the composition theorem to argue about privacy. This increases the variance of the noise by roughly k^2 with respect to only releasing a single function.

We can, however, do better by just using the Gaussian noise mechanism once. Define the function

$$f(X) = \begin{pmatrix} f_1(X) \\ \vdots \\ f_k(X) \end{pmatrix},$$

and observe that

$$\begin{aligned}
\Delta_2 f^2 &= \max_{X \sim X'} \|f(X) - f(X')\|_2^2 \\
&= \max_{X \sim X'} \sum_i \|f_i(X) - f_i(X')\|_2^2 \\
&\leq \sum_i \max_{X \sim X'} \|f_i(X) - f_i(X')\|_2^2 \\
&\leq kC^2.
\end{aligned}$$

Then we can release

$$f(X) + Z = \begin{pmatrix} f_1(X) + Z_1 \\ \vdots \\ f_k(X) + Z_k \end{pmatrix}$$

where each $Z_i \sim \mathcal{N}\left(0, \frac{kC^2}{\rho_{\varepsilon, \delta}} \cdot I\right)$, which is only k times the variance needed for releasing a single function.

However, the analysis via the composition theorem still had something going for it, because it still works even when the choice of the function f_i is determined by $f_1(X) + Z_1, \dots, f_{i-1}(X) + Z_{i-1}$. Can we get privacy with the improved variance above in this adaptive setting too? It turns that the answer is “yes”: in fact the privacy analysis of the Gaussian noise mechanism we did in the beginning of the course can be easily adapted to this setting. We have the following general result.

Theorem 1 *Suppose that we release $Y_1 = f_1(X) + Z_1, \dots, Y_k = f_k(X) + Z_k$ where $f_i : \mathcal{X}^n \rightarrow \mathbb{R}^d$ depends also on Y_1, \dots, Y_{i-1} . Suppose that each Z_i is independent and sampled from $\mathcal{N}\left(0, \frac{(\Delta_2 f_i)^2}{\rho_i} \cdot I\right)$. Then the output Y_1, \dots, Y_k satisfies (ε, δ) -DP for $\varepsilon = k\rho + \sqrt{2k\rho \ln(1/\delta)}$ and $\rho = \rho_1 + \dots + \rho_k$.*

A particular special case of Theorem 1 is that we can release any, potentially adaptively chosen, functions f_1, \dots, f_k , each of ℓ_2 -sensitivity at most C , by adding to $f_i(X)$ the noise $Z_i \sim \mathcal{N}\left(0, \frac{kC^2}{\rho_{\varepsilon, \delta}} \cdot I\right)$, and the final output is ε, δ -differentially private.

This is an instance of the advanced composition theorem: check the Dwork and Roth monograph for a more thorough discussion of it. Here we are stating it only for Gaussian noise, but, just like the simple composition theorem, a similar statement holds for adaptive composition of arbitrary (ε, δ) -differentially private mechanisms. The proof is again similar to the analysis of the Gaussian noise mechanism, but the parameters become worse, which is why we will stick to the Gaussian version.

5 Back to Noisy Gradient Descent

Let us now apply the Gaussian advanced composition theorem to the noisy gradient descent algorithm and determine how large the noise needs to be at each step. In this case the function f_1, f_2, \dots are simply the gradients $\nabla L_X(\theta_0), \nabla L_X(\theta_1)$, etc. To apply the composition theorem, we need to bound the sensitivity of these gradients, now seen as functions of X .

Suppose $\|\nabla \ell(\theta, (x, y))\|_2 \leq C$ for every $\theta \in \Theta$ and every $x \in \mathcal{X}, y \in \{-1, +1\}$, where again we take gradients with respect to θ .

Then, because $\nabla L_X(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla \ell(\theta, (x_i, y_i))$,

$$\begin{aligned} \Delta_2 \nabla L_X(\theta) &= \max_{X \sim X'} \|\nabla L_X(\theta) - \nabla L_{X'}(\theta)\|_2 \\ &= \max_{(x,y),(x',y') \in \mathcal{X} \times \{\pm 1\}} \frac{1}{n} \|\nabla \ell(\theta, (x, y)) - \nabla \ell(\theta, (x', y'))\|_2 \\ &\leq \max_{x \in \mathcal{X}, y \in \{\pm 1\}} \frac{2}{n} \|\nabla \ell(\theta, (x, y))\|_2 \leq \frac{2C}{n}. \end{aligned}$$

So, in private gradient descent, we can set $Z_t \sim \mathcal{N}\left(0, \frac{4TC^2}{n^2 \rho_{\varepsilon, \delta}} \cdot I\right)$, for each t in order to achieve (ε, δ) -differential privacy.

For logistic regression: with the loss $\ell(\theta, (x_i, y_i)) = \log(1 + e^{-y_i \langle \theta, x_i \rangle})$ we have

$$\nabla \log(1 + e^{-y_i \langle \theta, x_i \rangle}) = -\frac{1}{1 + e^{y_i \langle \theta, x_i \rangle}} y_i x_i \Rightarrow \|\nabla \log(1 + e^{-y_i \langle \theta, x_i \rangle})\|_2 \leq \|x_i\|_2 \leq \sqrt{d+1}.$$

So $Z_t \sim \mathcal{N}\left(0, \frac{4T(d+1)}{n^2 \rho_{\varepsilon, \delta}} \cdot I\right)$ is enough noise to achieve (ε, δ) -differential privacy.

6 Error analysis

Finally, we want to say that the noisy gradient descent algorithm, given sufficient data, does actually come close to the optimal parameter vector of the logistic regression problem. Luckily, gradient descent is an incredibly robust algorithm, and it's possible to still give convergence guarantees under a very general noise model.

Algorithms like the noisy gradient descent algorithm above are special cases of Stochastic Gradient Descent. In general, in stochastic gradient descent we take a step in a random direction, which ideally has expectation equal to minus the gradient, and not too large variance. The general algorithm is

```

 $\theta_0 \leftarrow 0$ 
for  $t = 1 \dots T - 1$  do
   $\tilde{\theta}_t \leftarrow \theta_{t-1} - \eta G_t$ 
   $\theta_t \leftarrow \tilde{\theta}_t / \max\{1, \|\theta_t\|_2 / R\}$ 
end for
output  $\frac{1}{T} \sum_{t=0}^{T-1} \theta_t$ 

```

Above G_t is a random variable, whose distribution may depend on $\theta_0, \dots, \theta_{t-1}$. In our case $G_t = \nabla L_X(\theta_{t-1}) + Z_t$. Another common variant of SGD is to set it equal to the gradient of a random point in the dataset, or the average of the gradients of several random points. This is often done to speed up the algorithm, since the most expensive step in gradient descent is to compute the gradient of the entire loss function.

In the next section we will prove the following general guarantee for SGD.

Theorem 2 *Let $L_X(\theta)$ be convex in θ for all X . Suppose $\mathbb{E}[G_t | \theta_{t-1}] = \nabla L_X(\theta_{t-1})$ and $\mathbb{E}\|G_t\|_2^2 \leq B^2$. For $\eta = \frac{R}{BT^{1/2}}$ we have $\mathbb{E}L_X\left(\frac{1}{T} \sum_{t=0}^{T-1} \theta_t\right) \leq \min_{\theta \in B_2^{d+1}(R)} L_X(\theta) + \frac{RB}{T^{1/2}}$.*

Using this theorem for private gradient descent, we have

$$G_t = \nabla L_X(\theta_{t-1}) + Z_t$$

$$\mathbb{E}\|G_t\|_2^2 = \|\nabla L_X(\theta_{t-1})\|_2^2 + \mathbb{E}\|Z_t\|_2^2 \leq C^2 + \frac{4TC^2(d+1)}{n^2\rho_{\varepsilon,\delta}} = C^2\left(1 + \frac{4T(d+1)}{n^2\rho_{\varepsilon,\delta}}\right)$$

This bounds B , and plugging the bound into Theorem 2, we get

$$\mathbb{E}L\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t, X\right) - \min_{\theta \in B_2^{d+1}(R)} L(\theta, X) \leq \frac{RB}{T^{1/2}} \leq \frac{RC}{T^{1/2}} \cdot \left(1 + \frac{4T(d+1)}{n^2\rho_{\varepsilon,\delta}}\right)^{1/2}$$

$$\leq \frac{RC}{T^{1/2}} + \frac{2RC\sqrt{d+1}}{n\sqrt{\rho_{\varepsilon,\delta}}}$$

Note that this error bound decomposes into two parts: one goes to 0 with T , and would be there even if we added no noise, i.e. just ran standard gradient descent. The second term in the error bound is due to the noise, and goes down with n , since the variance of our noise decreases with n as well. To bound the error by α , we set both terms to be less than $\frac{\alpha}{2}$ and we get the following setting of parameters

$$T \geq \frac{4R^2C^2}{\alpha^2}; \quad n \geq \frac{4RC\sqrt{d+1}}{\alpha\sqrt{\rho_{\varepsilon,\delta}}} \gtrsim \frac{RC\sqrt{d}\sqrt{\log(1/\delta)}}{\alpha\varepsilon}.$$

In the case of logistic regression, we can just plug in $C \leq \sqrt{d+1}$ in the bounds above and get

$$T \geq \frac{4R^2(d+1)}{\alpha^2}; \quad n \gtrsim \frac{Rd\sqrt{\log(1/\delta)}}{\alpha\varepsilon}.$$

7 Proof of Theorem 2

To be more concise, let's define $g(\theta) = L_X(\theta)$. We start with a basic fact about convex functions: the graph of a convex function always lies above any of its tangent hyperplanes (see Figure 2 for what this looks like in a single variable). This means that for all θ and θ' , we have

$$g(\theta') \geq g(\theta) + \langle \nabla g(\theta), \theta' - \theta \rangle \tag{1}$$

Stated equivalently, the local linear approximation to g at θ (on the right hand side) is always an underestimate with respect to the actual function.

Let $\theta^* = \arg \min_{\theta \in \Theta} g(\theta)$ be the optimal solution. Applying (1) to θ^* and θ_{t-1} , for any $1 \leq t \leq T$, we get

$$g(\theta^*) \geq g(\theta_{t-1}) + \langle \nabla g(\theta_{t-1}), \theta^* - \theta_{t-1} \rangle$$

or, equivalently,

$$g(\theta_{t-1}) - g(\theta^*) \leq \langle \nabla g(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle = \mathbb{E}[\langle G_t, \theta_{t-1} - \theta^* \rangle \mid \theta_{t-1}].$$

Taking expectations over θ_{t-1} , by the total law of expectation, we have

$$\mathbb{E}[g(\theta_{t-1}) - g(\theta^*)] \leq \mathbb{E}[\mathbb{E}[\langle G_t, \theta_{t-1} - \theta^* \rangle \mid \theta_{t-1}]] = \mathbb{E}[\langle G_t, \theta_{t-1} - \theta^* \rangle]$$

$$= \frac{1}{\eta} \mathbb{E}[\langle \theta_{t-1} - \tilde{\theta}_t, \theta_{t-1} - \theta^* \rangle] = (\star)$$

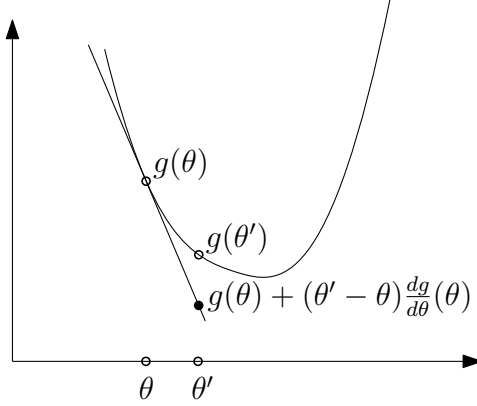


Figure 2: Illustration of (1)

Next, we use a nice little trick, known as polarization: for any two vectors x and y in \mathbb{R}^{d+1} , we have $\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle$, so we can express their dot product as $\langle x, y \rangle = \frac{1}{2}(\|x\|_2^2 + \|y\|_2^2 - \|x - y\|_2^2)$. Applying this to $x = \theta_{t-1} - \theta^*$ and $y = \theta_{t-1} - \tilde{\theta}_t$, on the right hand side above, we get

$$\begin{aligned}
(\star) &= \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \tilde{\theta}_t\|_2^2 + \|\theta_{t-1} - \theta^*\|_2^2 - \|\theta^* - \tilde{\theta}_t\|_2^2] \\
&= \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \theta^*\|_2^2 - \|\theta^* - \tilde{\theta}_t\|_2^2] + \frac{\eta}{2} \mathbb{E}[\|G_t\|_2^2] \\
&\leq \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \theta^*\|_2^2 - \|\theta^* - \tilde{\theta}_t\|_2^2] + \frac{\eta B^2}{2} = (\star\star).
\end{aligned}$$

Now we need one final inequality:

$$\|\theta^* - \theta_t\|_2 \leq \|\theta^* - \tilde{\theta}_t\|_2. \quad (2)$$

This follows from a simple geometric argument. When $\|\tilde{\theta}_t\|_2 \leq R$ the two sides above are equal. Otherwise, $\theta_t = R\tilde{\theta}_t/\|\tilde{\theta}_t\|_2$ is the closest point to $\tilde{\theta}_t$ inside the ball $B_2^{d+1}(R)$. This means that θ^* , $\tilde{\theta}_t$ and θ_t form a triangle with a non-acute (i.e. right or obtuse) angle at θ_t , as in Figure 3. Therefore, the right hand side in (2) equals the length of the side of the triangle opposite the non-acute angle, and the left hand side is the length of one of the other sides, which can be only smaller, by the cosine law.

Plugging (2) back into our calculations, we see that

$$(\star\star) \leq \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \theta^*\|_2^2 - \|\theta^* - \theta_t\|_2^2] + \frac{\eta B^2}{2}.$$

Putting everything together, we have shown that

$$\begin{aligned}
\mathbb{E}[g(\theta_{t-1}) - g(\theta^*)] &\leq \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \theta^*\|_2^2 - \|\theta^* - \theta_t\|_2^2] + \frac{\eta B^2}{2} \\
&= \frac{1}{2\eta} \mathbb{E}[\|\theta_{t-1} - \theta^*\|_2^2 - \|\theta_t - \theta^*\|_2^2] + \frac{\eta B^2}{2}
\end{aligned}$$

We have one of these inequalities for each $t \in \{1, \dots, T\}$, and if we average them, the right hand sides

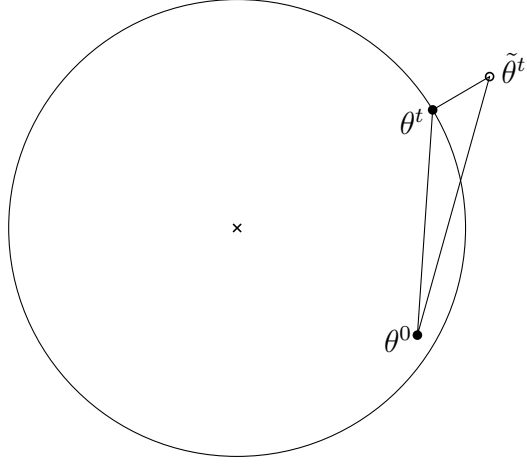


Figure 3: Projection back to the ball

telescope, and we get

$$\begin{aligned}
\mathbb{E} \left[g \left(\frac{1}{T} \sum_{t=0}^{T-1} \theta_t \right) - g(\theta^*) \right] &\leq \mathbb{E} \left[\frac{1}{T} \sum_{t=0}^{T-1} g(\theta_t) - g(\theta^*) \right] \\
&= \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T g(\theta_{t-1}) - g(\theta^*) \right] \\
&\leq \frac{1}{2T\eta} \mathbb{E} [\|\theta_0 - \theta^*\|_2^2 - \|\theta_T - \theta^*\|_2^2] + \frac{\eta B^2}{2} \\
&\leq \frac{R^2}{2T\eta} + \frac{\eta B^2}{2}.
\end{aligned}$$

The first inequality above follows from the convexity of g . The final inequality follows because $\theta_0 = 0$ and $\theta^* \in B_2^{d+1}(R)$, so $\|\theta_0 - \theta^*\|_2 \leq R$. Optimizing over the choice of η finishes the proof.