Introduction to Machine Learning (67577)  
Lecture 7

Shai Shalev-Shwartz

School of CS and Engineering,  
The Hebrew University of Jerusalem

Solving Convex Problems using SGD and RLM
1. Reminder: Convex learning problems

2. Learning Using Stochastic Gradient Descent

3. Learning Using Regularized Loss Minimization

4. Dimension vs. Norm bounds
   - Example application: Text categorization
Definition (Convex-Lipschitz-Bounded Learning Problem)

A learning problem, \((\mathcal{H}, Z, \ell)\), is called Convex-Lipschitz-Bounded, with parameters \(\rho, B\) if the following holds:

- The hypothesis class \(\mathcal{H}\) is a convex set and for all \(w \in \mathcal{H}\) we have \(\|w\| \leq B\).
- For all \(z \in Z\), the loss function, \(\ell(\cdot, z)\), is a convex and \(\rho\)-Lipschitz function.

Example:

- \(\mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}\)
- \(X = \{x \in \mathbb{R}^d : \|x\| \leq \rho\}\)
- \(Y = \mathbb{R}\)
- \(\ell(w, (x, y)) = |\langle w, x \rangle - y|\)
A learning problem, \((\mathcal{H}, Z, \ell)\), is called Convex-Lipschitz-Bounded, with parameters \(\rho, B\) if the following holds:

- The hypothesis class \(\mathcal{H}\) is a convex set and for all \(w \in \mathcal{H}\) we have \(\|w\| \leq B\).
- For all \(z \in Z\), the loss function, \(\ell(\cdot, z)\), is a convex and \(\rho\)-Lipschitz function.

Example:

- \(\mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}\)
- \(\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq \rho\}, \mathcal{Y} = \mathbb{R}\),
- \(\ell(w, (x, y)) = |\langle w, x \rangle - y|\)
Definition (Convex-Smooth-Bounded Learning Problem)

A learning problem, \((\mathcal{H}, Z, \ell)\), is called Convex-Smooth-Bounded, with parameters \(\beta, B\) if the following holds:

- The hypothesis class \(\mathcal{H}\) is a convex set and for all \(w \in \mathcal{H}\) we have \(\|w\| \leq B\).
- For all \(z \in Z\), the loss function, \(\ell(\cdot, z)\), is a convex, non-negative, and \(\beta\)-smooth function.

Example:

\[ \mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\} \]
\[ X = \{x \in \mathbb{R}^d : \|x\| \leq \beta/2\} \]
\[ Y = \mathbb{R} \]
\[ \ell(w, (x, y)) = (\langle w, x \rangle - y)^2 \]
Convex-Smooth-bounded learning problem

Definition (Convex-Smooth-Bounded Learning Problem)

A learning problem, \((\mathcal{H}, Z, \ell)\), is called Convex-Smooth-Bounded, with parameters \(\beta, B\) if the following holds:

- The hypothesis class \(\mathcal{H}\) is a convex set and for all \(w \in \mathcal{H}\) we have \(\|w\| \leq B\).
- For all \(z \in Z\), the loss function, \(\ell(\cdot, z)\), is a convex, non-negative, and \(\beta\)-smooth function.

Example:

- \(\mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}\)
- \(\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq \beta/2\}\), \(\mathcal{Y} = \mathbb{R}\),
- \(\ell(w, (x, y)) = (\langle w, x \rangle - y)^2\)
Outline

1. Reminder: Convex learning problems

2. Learning Using Stochastic Gradient Descent

3. Learning Using Regularized Loss Minimization

4. Dimension vs. Norm bounds
   - Example application: Text categorization
Consider a learning problem.
Consider a learning problem.

Recall: our goal is to (probably approximately) solve:

$$\min_{w \in \mathcal{H}} L_{\mathcal{D}}(w) \text{ where } L_{\mathcal{D}}(w) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(w, z)]$$
Learning Using Stochastic Gradient Descent

- Consider a learning problem.
- Recall: our goal is to (probably approximately) solve:
  \[
  \min_{w \in \mathcal{H}} L_D(w) \quad \text{where} \quad L_D(w) = \mathbb{E}_{z \sim D} [\ell(w, z)]
  \]
- So far, learning was based on the empirical risk, \( L_S(w) \)
Consider a learning problem.

Recall: our goal is to (probably approximately) solve:

$$\min_{w \in H} L_D(w) \quad \text{where} \quad L_D(w) = \mathbb{E}_{z \sim D}[\ell(w, z)]$$

So far, learning was based on the empirical risk, $L_S(w)$

We now consider directly minimizing $L_D(w)$
\[ \min_{w \in \mathcal{H}} L_D(w) \quad \text{where} \quad L_D(w) = \mathbb{E}_{z \sim D}[\ell(w, z)] \]

Recall the gradient descent method in which we initialize \( w^{(1)} = 0 \) and update \( w^{(t+1)} = w^{(t)} - \eta \nabla L_D(w) \).
Stochastic Gradient Descent

\[
\min_{w \in \mathcal{H}} L_D(w) \quad \text{where} \quad L_D(w) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(w, z)]
\]

- Recall the gradient descent method in which we initialize \( w^{(1)} = 0 \) and update \( w^{(t+1)} = w^{(t)} - \eta \nabla L_D(w) \)
- Observe: \( \nabla L_D(w) = \mathbb{E}_{z \sim \mathcal{D}}[\nabla \ell(w, z)] \)
Stochastic Gradient Descent

\[
\min_{w \in \mathcal{H}} L_D(w) \text{ where } L_D(w) = \mathbb{E}_{z \sim D} [\ell(w, z)]
\]

- Recall the gradient descent method in which we initialize \( w^{(1)} = 0 \) and update \( w^{(t+1)} = w^{(t)} - \eta \nabla L_D(w) \)
- Observe: \( \nabla L_D(w) = \mathbb{E}_{z \sim D} [\nabla \ell(w, z)] \)
- We can’t calculate \( \nabla L_D(w) \) because we don’t know \( D \)
Stochastic Gradient Descent

\[
\min_{\mathbf{w} \in \mathcal{H}} L_D(\mathbf{w}) \quad \text{where} \quad L_D(\mathbf{w}) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\mathbf{w}, z)]
\]

- Recall the gradient descent method in which we initialize \( \mathbf{w}^{(1)} = 0 \) and update \( \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla L_D(\mathbf{w}) \)
- Observe: \( \nabla L_D(\mathbf{w}) = \mathbb{E}_{z \sim \mathcal{D}}[\nabla \ell(\mathbf{w}, z)] \)
- We can’t calculate \( \nabla L_D(\mathbf{w}) \) because we don’t know \( \mathcal{D} \)
- But we can estimate it by \( \nabla \ell(\mathbf{w}, \mathbf{z}) \) for \( \mathbf{z} \sim \mathcal{D} \)
Stochastic Gradient Descent

$$\min_{w \in \mathcal{H}} L_{D}(w) \quad \text{where} \quad L_{D}(w) = \mathbb{E}_{z \sim D}[\ell(w, z)]$$

- Recall the gradient descent method in which we initialize $w^{(1)} = 0$ and update $w^{(t+1)} = w^{(t)} - \eta \nabla L_{D}(w)$
- Observe: $\nabla L_{D}(w) = \mathbb{E}_{z \sim D}[\nabla \ell(w, z)]$
- We can’t calculate $\nabla L_{D}(w)$ because we don’t know $D$
- But we can estimate it by $\nabla \ell(w, z)$ for $z \sim D$
- If we take a step based on the direction $v = \nabla \ell(w, z)$ then in expectation we’re moving in the right direction
Stochastic Gradient Descent

\[
\min_{w \in \mathcal{H}} L_D(w) \quad \text{where} \quad L_D(w) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(w, z)]
\]

- Recall the gradient descent method in which we initialize \( w^{(1)} = 0 \) and update \( w^{(t+1)} = w^{(t)} - \eta \nabla L_D(w) \)
- Observe: \( \nabla L_D(w) = \mathbb{E}_{z \sim \mathcal{D}}[\nabla \ell(w, z)] \)
- We can’t calculate \( \nabla L_D(w) \) because we don’t know \( \mathcal{D} \)
- But we can estimate it by \( \nabla \ell(w, z) \) for \( z \sim \mathcal{D} \)
- If we take a step based on the direction \( v = \nabla \ell(w, z) \) then in expectation we’re moving in the right direction
- In other words, \( v \) is an unbiased estimate of the gradient
Stochastic Gradient Descent

\[ \min_{w \in \mathcal{H}} L_D(w) \text{ where } L_D(w) = \mathbb{E}_{z \sim D}[\ell(w, z)] \]

- Recall the gradient descent method in which we initialize \( w^{(1)} = 0 \) and update \( w^{(t+1)} = w^{(t)} - \eta \nabla L_D(w) \)
- Observe: \( \nabla L_D(w) = \mathbb{E}_{z \sim D}[\nabla \ell(w, z)] \)
- We can’t calculate \( \nabla L_D(w) \) because we don’t know \( D \)
- But we can estimate it by \( \nabla \ell(w, z) \) for \( z \sim D \)
- If we take a step based on the direction \( v = \nabla \ell(w, z) \) then in expectation we’re moving in the right direction
- In other words, \( v \) is an unbiased estimate of the gradient
- We’ll show that this is good enough
Stochastic Gradient Descent

- **initialize:** $w^{(1)} = 0$
- **for** $t = 1, 2, \ldots, T$
  - choose $z_t \sim \mathcal{D}$
  - let $v_t \in \partial \ell(w^{(t)}, z_t)$ update $w^{(t+1)} = w^{(t)} - \eta v_t$
- **output** $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$
Stochastic Gradient Descent

- **initialize**: $w^{(1)} = 0$
- **for** $t = 1, 2, \ldots, T$
  - choose $z_t \sim D$
  - let $v_t \in \partial \ell(w^{(t)}, z_t)$ update $w^{(t+1)} = w^{(t)} - \eta v_t$
- **output** $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$
Analyzing SGD for convex-Lipschitz-bounded

By algebraic manipulations, for any sequence of $v_1, \ldots, v_T$, and any $w^*$,

$$
\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{\| w^{(1)} - w^* \|^2 - \| w^{(T+1)} - w^* \|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \| v_t \|^2
$$

Assume that $\| v_t \| \leq \rho$ for all $t$ and that $\| w^* \| \leq B$ we obtain

$$
\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq B\rho \sqrt{T}.
$$
Analyzing SGD for convex-Lipschitz-bounded

By algebraic manipulations, for any sequence of $v_1, \ldots, v_T$, and any $w^*$,

$$
\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{\| w^{(1)} - w^* \|^2 - \| w^{(T+1)} - w^* \|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \| v_t \|^2
$$

Assume that $\| v_t \| \leq \rho$ for all $t$ and that $\| w^* \| \leq B$ we obtain

$$
\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{B^2}{2\eta} + \frac{\eta \rho^2 T}{2}
$$
Analyzing SGD for convex-Lipschitz-bounded

By algebraic manipulations, for any sequence of $v_1, \ldots, v_T$, and any $w^*$,

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{\|w^{(1)} - w^*\|^2 - \|w^{(T+1)} - w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

Assume that $\|v_t\| \leq \rho$ for all $t$ and that $\|w^*\| \leq B$ we obtain

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{B^2}{2\eta} + \frac{\eta \rho^2 T}{2}$$

In particular, for $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ we get

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq B \rho \sqrt{T}$$
Analyzing SGD for convex-Lipschitz-bounded

Taking expectation of both sides w.r.t. the randomness of choosing $z_1, \ldots, z_T$ we obtain:

$$\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \right] \leq B \rho \sqrt{T}.$$
Analyzing SGD for convex-Lipschitz-bounded

Taking expectation of both sides w.r.t. the randomness of choosing $z_1, \ldots, z_T$ we obtain:

$$
\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \right] \leq B \rho \sqrt{T}.
$$

The law of total expectation: for every two random variables $\alpha, \beta$, and a function $g$, $\mathbb{E}_{\alpha}[g(\alpha)] = \mathbb{E}_{\beta} \mathbb{E}_{\alpha}[g(\alpha)|\beta]$. 
Analyzing SGD for convex-Lipschitz-bounded

Taking expectation of both sides w.r.t. the randomness of choosing $z_1,\ldots,z_T$ we obtain:

$$
\mathbb{E}_{z_1,\ldots,z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \right] \leq B \rho \sqrt{T}.
$$

The law of total expectation: for every two random variables $\alpha, \beta$, and a function $g$, $\mathbb{E}_\alpha[g(\alpha)] = \mathbb{E}_\beta \mathbb{E}_\alpha[g(\alpha)|\beta]$. Therefore

$$
\mathbb{E}_{z_1,\ldots,z_T} \left[ \langle w^{(t)} - w^*, v_t \rangle \right] = \mathbb{E}_{z_1,\ldots,z_{t-1}} \mathbb{E}_{z_1,\ldots,z_T} \left[ \langle w^{(t)} - w^*, v_t \rangle | z_1,\ldots,z_{t-1} \right].
$$
Analyzing SGD for convex-Lipschitz-bounded

Taking expectation of both sides w.r.t. the randomness of choosing $z_1, \ldots, z_T$ we obtain:

$$
\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \right] \leq B \rho \sqrt{T}.
$$

The law of total expectation: for every two random variables $\alpha, \beta$, and a function $g$, $\mathbb{E}_\alpha [g(\alpha)] = \mathbb{E}_\beta \mathbb{E}_\alpha [g(\alpha) | \beta]$. Therefore

$$
\mathbb{E}_{z_1, \ldots, z_T} [\langle w^{(t)} - w^*, v_t \rangle] = \mathbb{E}_{z_1, \ldots, z_{t-1}} \mathbb{E}_{z_1, \ldots, z_T} [\langle w^{(t)} - w^*, v_t \rangle | z_1, \ldots, z_{t-1}].
$$

Once we know $z_1, \ldots, z_{t-1}$ the value of $w^{(t)}$ is not random, hence,

$$
\mathbb{E}_{z_1, \ldots, z_T} [\langle w^{(t)} - w^*, v_t \rangle | z_1, \ldots, z_{t-1}] = \langle w^{(t)} - w^*, \mathbb{E}_{z_t} [\nabla \ell(w^{(t)}, z_t)] \rangle
$$

$$
= \langle w^{(t)} - w^*, \nabla L_D(w^{(t)}) \rangle
$$
Analyzing SGD for convex-Lipschitz-bounded

We got:

$$\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla L_D(w^{(t)}) \rangle \right] \leq B \rho \sqrt{T}$$
Analyzing SGD for convex-Lipschitz-bounded

We got:

\[
\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla L_D(w^{(t)}) \rangle \right] \leq B \rho \sqrt{T}
\]

By convexity, this means

\[
\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} (L_D(w^{(t)}) - L_D(w^*)) \right] \leq B \rho \sqrt{T}
\]
Analyzing SGD for convex-Lipschitz-bounded

We got:

$$\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla L_D(w^{(t)}) \rangle \right] \leq B \rho \sqrt{T}$$

By convexity, this means

$$\mathbb{E}_{z_1, \ldots, z_T} \left[ \sum_{t=1}^{T} (L_D(w^{(t)}) - L_D(w^*)) \right] \leq B \rho \sqrt{T}$$

Dividing by $T$ and using convexity again,

$$\mathbb{E}_{z_1, \ldots, z_T} \left[ L_D \left( \frac{1}{T} \sum_{t=1}^{T} w^{(t)} \right) \right] \leq L_D(w^*) + \frac{B \rho}{\sqrt{T}}$$
Corollary

Consider a convex-Lipschitz-bounded learning problem with parameters $\rho, B$. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_D(w)$ with a number of iterations (i.e., number of examples) $T \geq \frac{B^2 \rho^2}{\epsilon^2}$

and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies:

$$E[L_D(\bar{w})] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon .$$
Corollary

Consider a convex-Lipschitz-bounded learning problem with parameters $\rho, B$. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_D(w)$ with a number of iterations (i.e., number of examples)

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies:

$$E[L_D(\overline{w})] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon.$$

- Remark: Can obtain high probability bound using “boosting the confidence” (Lecture 4)
Convex-smooth-bounded problems

Similar result holds for smooth problems:

**Corollary**

Consider a convex-smooth-bounded learning problem with parameters $\beta, B$. Assume in addition that $\ell(0, z) \leq 1$ for all $z \in Z$. For every $\epsilon > 0$, set $\eta = \frac{1}{\beta (1 + 3/\epsilon)}$. Then, running SGD with $T \geq 12B^2 \beta / \epsilon^2$ yields

$$
\mathbb{E}[L_D(\bar{w})] \leq \min_{w \in H} L_D(w) + \epsilon .
$$
Outline

1. Reminder: Convex learning problems

2. Learning Using Stochastic Gradient Descent

3. Learning Using Regularized Loss Minimization

4. Dimension vs. Norm bounds
   - Example application: Text categorization
Regularized Loss Minimization (RLM)

Given a regularization function $R : \mathbb{R}^d \to \mathbb{R}$, the RLM rule is:

$$A(S) = \arg\min_w (L_S(w) + R(w))$$.
Regularized Loss Minimization (RLM)

Given a regularization function $R : \mathbb{R}^d \rightarrow \mathbb{R}$, the RLM rule is:

$$A(S) = \arg\min_w (L_S(w) + R(w)) .$$

We will focus on Tikhonov regularization

$$A(S) = \arg\min_w (L_S(w) + \lambda \|w\|^2) .$$
Why to regularize?

- **Similar to MDL**: specify “prior belief” in hypotheses. We bias ourselves toward “short” vectors.
- **Stabilizer**: we’ll show that Tikhonov regularization makes the learner stable w.r.t. small perturbation of the training set, which in turn leads to generalization.
Stability

- **Informally**: an algorithm $A$ is stable if a small change of its input $S$ will lead to a small change of its output hypothesis.
Stability

- **Informally**: an algorithm $A$ is stable if a small change of its input $S$ will lead to a small change of its output hypothesis.
- Need to specify what is “small change of input” and what is “small change of output”. 
Stability

- Replace one sample: given $S = (z_1, \ldots, z_m)$ and an additional example $z'$, let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$
Stability

- Replace one sample: given $S = (z_1, \ldots, z_m)$ and an additional example $z'$, let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$

Definition (on-average-replace-one-stable)

Let $\epsilon : \mathbb{N} \rightarrow \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm $A$ is on-average-replace-one-stable with rate $\epsilon(m)$ if for every distribution $D$

$$
\mathbb{E}_{(S,z') \sim D^{m+1}, i \sim U(m)} [\ell(A(S^{(i)}, z_i)) - \ell(A(S), z_i)] \leq \epsilon(m).
$$
Stable rules do not overfit

**Theorem**

If $A$ is on-average-replace-one-stable with rate $\epsilon(m)$ then

$$\mathbb{E}_{S \sim \mathcal{D}_m}[L_D(A(S)) - L_S(A(S))] \leq \epsilon(m).$$
Stable rules do not overfit

**Theorem**

If $A$ is on-average-replace-one-stable with rate $\epsilon(m)$ then

$$\mathbb{E}_{S \sim \mathcal{D}^m} [L_D(A(S)) - L_S(A(S))] \leq \epsilon(m).$$

**Proof.**

Since $S$ and $z'$ are both drawn i.i.d. from $\mathcal{D}$, we have that for every $i$,

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_{S,z'}[\ell(A(S), z')] = \mathbb{E}_{S,z'}[\ell(A(S^{(i)}), z_i)].$$

On the other hand, we can write

$$\mathbb{E}_S[L_S(A(S))] = \mathbb{E}_{S,i}[\ell(A(S), z_i)].$$

The proof follows from the definition of stability.
Assume that the loss function is convex and $\rho$-Lipschitz. Then, the RLM rule with the regularizer $\lambda \|w\|^2$ is on-average-replace-one-stable with rate $\frac{2 \rho^2}{\lambda m}$. It follows that

$$\mathbb{E}_{S \sim D_m}[L_D(A(S)) - L_S(A(S))] \leq \frac{2 \rho^2}{\lambda m}.$$ 

Similarly, for convex, $\beta$-smooth, and non-negative, loss the rate is $\frac{48 \beta C}{\lambda m}$, where $C$ is an upper bound on $\max_z \ell(0, z)$. 

Tikhonov Regularization as Stabilizer

**Theorem**

Assume that the loss function is convex and $\rho$-Lipschitz. Then, the RLM rule with the regularizer $\lambda \|w\|^2$ is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$. It follows that

$$\mathbb{E}_{S \sim D_m}[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}.$$

Similarly, for convex, $\beta$-smooth, and non-negative, loss the rate is $\frac{48\beta C}{\lambda m}$, where $C$ is an upper bound on $\max_z \ell(0, z)$.

The proof relies on the notion of strong convexity and can be found in the book.
Observe:

\[
\mathbb{E}_S[\mathcal{L}_D(A(S))] = \mathbb{E}_S[\mathcal{L}_S(A(S))] + \mathbb{E}_S[\mathcal{L}_D(A(S)) - \mathcal{L}_S(A(S))].
\]
The Fitting-Stability Tradeoff

Observe:

\[ \mathbb{E}_S[L_D(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_D(A(S)) - L_S(A(S))] . \]

- The first term is how good \( A \) fits the training set.
- The 2nd term is the overfitting, and is bounded by the stability of \( A \).
The Fitting-Stability Tradeoff

Observe:

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_D(A(S)) - L_S(A(S))] .$$

- The first term is how good $A$ fits the training set
- The 2nd term is the overfitting, and is bounded by the stability of $A$
- $\lambda$ controls the tradeoff between the two terms
Let $A$ be the RLM rule
The Fitting-Stability Tradeoff

- Let $A$ be the RLM rule
- We saw (for convex-Lipschitz losses)

$$
\mathbb{E}_S[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}
$$
Let $A$ be the RLM rule

We saw (for convex-Lipschitz losses)

$$
\mathbb{E}_S[L_D(A(S)) - L_S(A(S)) \leq \frac{2\rho^2}{\lambda m}
$$

Fix some arbitrary vector $w^*$, then:

$$
L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(w^*) + \lambda\|w^*\|^2.
$$
The Fitting-Stability Tradeoff

- Let $A$ be the RLM rule
- We saw (for convex-Lipschitz losses)
  \[
  \mathbb{E}_S[L_D(A(S)) - L_S(A(S))] \leq \frac{2 \rho^2}{\lambda m}
  \]
- Fix some arbitrary vector $w^*$, then:
  \[
  L_S(A(S)) \leq L_S(A(S)) + \lambda \|A(S)\|^2 \leq L_S(w^*) + \lambda \|w^*\|^2.
  \]
- Taking expectation of both sides with respect to $S$ and noting that \(\mathbb{E}_S[L_S(w^*)] = L_D(w^*)\), we obtain that
  \[
  \mathbb{E}_S[L_S(A(S))] \leq L_D(w^*) + \lambda \|w^*\|^2.
  \]
The Fitting-Stability Tradeoff

- Let $A$ be the RLM rule
- We saw (for convex-Lipschitz losses)

$$\mathbb{E}_S[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}$$

- Fix some arbitrary vector $w^*$, then:

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(w^*) + \lambda\|w^*\|^2.$$  

- Taking expectation of both sides with respect to $S$ and noting that $\mathbb{E}_S[L_S(w^*)] = L_D(w^*)$, we obtain that

$$\mathbb{E}_S[L_S(A(S))] \leq L_D(w^*) + \lambda\|w^*\|^2.$$  

- Therefore:

$$\mathbb{E}_S[L_D(A(S))] \leq L_D(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m}$$
The Regularization Path

The RLM rule as a function of $\lambda$ is $\mathbf{w}(\lambda) = \arg\min_{\mathbf{w}} L_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2$
The RLM rule as a function of $\lambda$ is $w(\lambda) = \arg\min_w L_S(w) + \lambda\|w\|^2$
Can be seen as a pareto objective: minimize both $L_S(w)$ and $\|w\|^2$
The RLM rule as a function of $\lambda$ is $w(\lambda) = \arg\min_w L_S(w) + \lambda \|w\|^2$
Can be seen as a pareto objective: minimize both $L_S(w)$ and $\|w\|^2$
How to choose $\lambda$?

- **Bound minimization**: choose $\lambda$ according to the bound on $L_D(w)$ usually far from optimal as the bound is worst case.

- **Validation**: calculate several pareto optimal points on the regularization path (by varying $\lambda$) and use validation set to choose the best one.
Outline

1. Reminder: Convex learning problems
2. Learning Using Stochastic Gradient Descent
3. Learning Using Regularized Loss Minimization
4. Dimension vs. Norm bounds
   - Example application: Text categorization
Previously in the course, when we learnt $d$ parameters the sample complexity grew with $d$. Here, we learn $d$ parameters but the sample complexity depends on the norm of $\|w^*\|$ and on the Lipschitzness/smoothness, rather than on $d$. Which approach is better depends on the properties of the distribution.
Signs all encouraging for Phelps in comeback. He did not win any gold medals or set any world records but Michael Phelps ticked all the boxes he needed in his comeback to competitive swimming.
Signs all encouraging for Phelps in comeback. He did not win any gold medals or set any world records but Michael Phelps ticked all the boxes he needed in his comeback to competitive swimming.
Document categorization

- Let $\mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1 \}$
Document categorization

- Let $\mathcal{X} = \{x \in \{0, 1\}^d : \|x\|^2 \leq R^2, x_d = 1\}$
- Think on $x \in \mathcal{X}$ as a text document represented as a bag of words:
Document categorization

- Let \( \mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1 \} \)

- Think on \( \mathbf{x} \in \mathcal{X} \) as a text document represented as a bag of words:
  - At most \( R^2 - 1 \) words in each document

- \( \mathcal{Y} = \{ \pm 1 \} \) (e.g., the document is about sport or not)

- Linear classifiers
  - \( \mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \)

  Intuitively:
  - \( w_i \) is large (positive) for words indicative to sport while \( w_i \) is small (negative) for words indicative to non-sport

- Hinge-loss:
  \[
  \ell(w, (x, y)) = \max(1 - y\langle w, x \rangle, 0)
  \]
Document categorization

- Let $\mathcal{X} = \{x \in \{0, 1\}^d : \|x\|^2 \leq R^2, x_d = 1\}$
- Think on $x \in \mathcal{X}$ as a text document represented as a \textbf{bag of words}:
  - At most $R^2 - 1$ words in each document
  - $d - 1$ is the size of the dictionary
Let $\mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1\}$

Think on $\mathbf{x} \in \mathcal{X}$ as a text document represented as a bag of words:

- At most $R^2 - 1$ words in each document
- $d - 1$ is the size of the dictionary
- Last coordinate is the bias
Document categorization

- Let $\mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1 \}$
- Think on $\mathbf{x} \in \mathcal{X}$ as a text document represented as a bag of words:
  - At most $R^2 - 1$ words in each document
  - $d - 1$ is the size of the dictionary
  - Last coordinate is the bias
- Let $\mathcal{Y} = \{ \pm 1 \}$ (e.g., the document is about sport or not)
Document categorization

- Let $\mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1 \}$
- Think on $\mathbf{x} \in \mathcal{X}$ as a text document represented as a bag of words:
  - At most $R^2 - 1$ words in each document
  - $d - 1$ is the size of the dictionary
  - Last coordinate is the bias
- Let $\mathcal{Y} = \{ \pm 1 \}$ (e.g., the document is about sport or not)
- Linear classifiers $\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$
Document categorization

- Let $\mathcal{X} = \{x \in \{0, 1\}^d : \|x\|^2 \leq R^2, x_d = 1\}$
- Think on $x \in \mathcal{X}$ as a text document represented as a bag of words:
  - At most $R^2 - 1$ words in each document
  - $d - 1$ is the size of the dictionary
  - Last coordinate is the bias
- Let $\mathcal{Y} = \{\pm 1\}$ (e.g., the document is about sport or not)
- Linear classifiers $x \mapsto \text{sign}(\langle w, x \rangle)$
- Intuitively: $w_i$ is large (positive) for words indicative to sport while $w_i$ is small (negative) for words indicative to non-sport
Document categorization

- Let \( \mathcal{X} = \{ \mathbf{x} \in \{0, 1\}^d : \|\mathbf{x}\|^2 \leq R^2, x_d = 1 \} \)
- Think on \( \mathbf{x} \in \mathcal{X} \) as a text document represented as a bag of words:
  - At most \( R^2 - 1 \) words in each document
  - \( d - 1 \) is the size of the dictionary
  - Last coordinate is the bias
- Let \( \mathcal{Y} = \{ \pm 1 \} \) (e.g., the document is about sport or not)
- Linear classifiers \( \mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \)
- Intuitively: \( w_i \) is large (positive) for words indicative to sport while \( w_i \) is small (negative) for words indicative to non-sport
- Hinge-loss: \( \ell(w, (x, y)) = [1 - y \langle w, x \rangle]_+ \)
VC dimension is $d$, but $d$ can be extremely large (number of words in English)
VC dimension is $d$, but $d$ can be extremely large (number of words in English)

- Loss function is convex and $R$ Lipschitz
VC dimension is $d$, but $d$ can be extremely large (number of words in English)

Loss function is convex and $R$ Lipschitz

Assume that the number of relevant words is small, and their weights is not too large, then there is a $\mathbf{w}^*$ with small norm and small $L_D(\mathbf{w}^*)$
VC dimension is $d$, but $d$ can be extremely large (number of words in English)

Loss function is convex and $R$ Lipschitz

Assume that the number of relevant words is small, and their weights is not too large, then there is a $w^*$ with small norm and small $L_D(w^*)$

Then, can learn it with sample complexity that depends on $R^2\|w^*\|^2$, and does not depend on $d$ at all!
VC dimension is $d$, but $d$ can be extremely large (number of words in English)

Loss function is convex and $R$ Lipschitz

Assume that the number of relevant words is small, and their weights is not too large, then there is a $w^*$ with small norm and small $L_D(w^*)$

Then, can learn it with sample complexity that depends on $R^2 \|w^*\|^2$, and does not depend on $d$ at all!

But, there are of course opposite cases, in which $d$ is much smaller than $R^2 \|w^*\|^2$
Learning convex learning problems using SGD
Learning convex learning problems using RLM
The regularization path
Dimension vs. Norm