1 Chapter 12: Convex Learning Problems

1.1 Convexity

Definition 1.1 (Convex Set)**.** A set C in a vector space is convex if for any $\boldsymbol{v}, \boldsymbol{w} \in C$, the line segment between $\boldsymbol{v}$ and $\boldsymbol{w}$ is contained in C. That is, for any $\alpha \in [0, 1]$, we have

$$\alpha \boldsymbol{v} + (1 - \alpha)\boldsymbol{w} \in C.$$

Example 2. Some convex shapes: $\mathbb{R}^d$, $\{\mathbf{0}\}$, (linear) cones.

Now, we will extend this notion to functions.

Definition 2.1 (Epigraph)**.** The *epigraph* of a function $f : X \to \mathbb{R}$ is the set

$$\text{epigraph}(f) = \{(\boldsymbol{x}, \beta) \mid f(\boldsymbol{x}) \leq \beta, \boldsymbol{x} \in X\}.$$

Definition 2.2 (Convex function)**.** A function f is convex if epigraph(f) is a convex set.

If we “plug-in” our definitions, we can get a more explicit form:

Definition 2.3 (Convex function, expanded)**.** Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\boldsymbol{v}, \boldsymbol{w} \in C$ and $\alpha \in [0, 1]$ we have

$$f(\alpha \boldsymbol{v} + (1 - \alpha)\boldsymbol{w}) \in \alpha f(\boldsymbol{v}) + (1 - \alpha) f(\boldsymbol{w}).$$

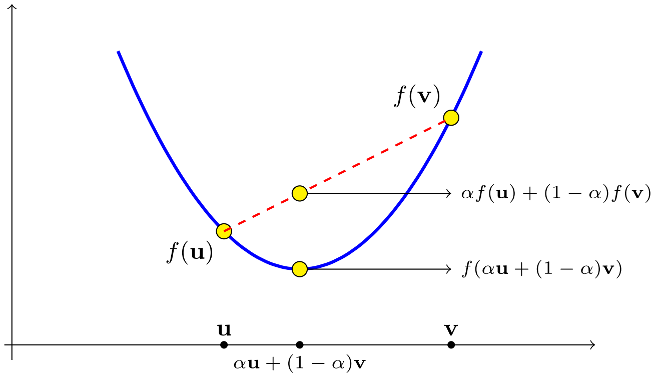
$f(\mathbf{v})$
$\alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$
$f(\mathbf{u})$
$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v})$
$\mathbf{u}$
$\mathbf{v}$
$\alpha\mathbf{u} + (1-\alpha)\mathbf{v}$

Figure 1: Graphical depiction of Proposition 2.3

If a convex function is also differentiable, then we have a nice property that it lies above its tangent plane:

Proposition 2.4. If f is convex and differentiable, then for all $\boldsymbol{v}, \boldsymbol{w}$, we have

$$f(\boldsymbol{v}) \geq \underbrace{f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle}_{\text{tangent plane}}$$

where $\nabla f(\boldsymbol{w}) = \left(\frac{\partial f(\boldsymbol{w})}{\partial w_1}, \ldots, \frac{\partial f(\boldsymbol{w})}{\partial w_d} \right)$.

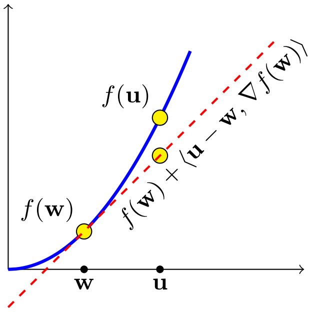
$f(\mathbf{u})$
$f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle$
$f(\mathbf{w})$
$\mathbf{w}$
$\mathbf{u}$

Figure 2: Graphical depiction of Proposition 2.4

Motivation 3. Convexity is a *local* property that gives us *global* structure.

3.1 Lipschitz continuity

We will be interested in bounding the sensitivity of functions.

Definition 3.1 (Lipschitzness)**.** Let $C \subseteq \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for every $\boldsymbol{w}_1, \boldsymbol{w}_2 \in C$ we have that

$$\|f(\boldsymbol{w}_1) - f(\boldsymbol{w}_2)\| \leq \rho \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|.$$

In other words, as we change the input to the function, the output changes at most ρ times as quickly. We call ρ the *Lipschitz constant* of f over C.

Example 4. Here are some Lipschitz (and non-Lipschitz) functions:

1. $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$.
2. $f(x) = x^2$ is not Lipschitz over $\mathbb{R}$ but it is 2-Lipschitz over $[-1, 1]$.
3. $f(x) = \sum_{n=1}^{\infty} \frac{2^{-n}}{n\pi} |\sin(n\pi x)|$ is 1-Lipschitz but not differentiable at any rational.

Proposition 4.1. Lipschitz functions are continuous and differentiable almost everywhere.

4.1 Smoothness (revisit)

Here is a definition of *smoothness*:

Definition 4.2 (Smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz.

You may have seen other definitions of smoothness (so don't get them mixed up!) This one is the most common in optimization.

Definition 4.3 (Smoothness, expanded)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if for all $\boldsymbol{v}, \boldsymbol{w}$, we have

$$\|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{w})\| \leq \rho \|\boldsymbol{v} - \boldsymbol{w}\|$$

In other words smooth functions have gradients that cannot change too fast.

Example 5. $f(x) = x^2$ is 2-smooth, but $f(x) = x^3$ is not smooth over $\mathbb{R}$.

6 Regularized Loss Minimization

6.1 Regularization

We will study a new learning paradigm: Regularized Loss Minimization and show that convex-Lipschitz-bounded, and convex-smooth-bounded families of learning problems are learnable. The key insight is that regularizers make learning algorithms more stable.

Definition 6.1 (Regularized Loss Minimization)**.**

$$A(S) = \operatorname*{argmin}_{\boldsymbol{w}}(L_S(\boldsymbol{w}) + R(\boldsymbol{w}))$$

The idea is that the regularizer R measures the complexity of the hypothesis. We are going to focus on *Tikhonov regularization*: $R(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2$ for $\lambda > 0$, which is also known as "ℓ^2 regularization".

Example 7. Applying Tikhonov regularization to logistic regression

$$A(X, \boldsymbol{y}) = \operatorname*{argmin}_{w \in \mathbb{R}^d}\left(\frac{1}{m}\sum_{i=1}^{m}\log\bigl(1 + \exp(-y\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\bigr) + \lambda\|\boldsymbol{w}\|^2\right).$$

Remark 7.1. Logistic regression without regularization is (strictly) convex. There is no closed form for the mimimizer this time, but there are efficient solvers for this problem!

Our goal is to show the following:

Theorem 1 (Regularized logistic regression test risk bound)**.** Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times [-1, 1]$, where $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$. Let $\mathcal{H} = \{\boldsymbol{w} \in \mathbb{R}^d \mid \|w\| \leq B\}$. For any $\epsilon \in (0, 1)$, let $m \geq 8B^2/\epsilon^2$. Then, applying regularized logistic regression with parameter $\lambda = \epsilon/(2B^2)$ satisfies

$$\mathbb{E}_{S \in D^m}[L_{\mathcal{D}}(A_{\mathrm{RLM}}(S))] \leq \min_{\boldsymbol{w} \in \mathcal{H}} L_{\mathcal{D}}(\boldsymbol{w}) + \epsilon.$$

Remark 7.2. Note:

1. Both $\mathcal{X}$ and $\mathcal{H}$ are bounded. We will learn that this is important...
2. Similarly, the regularization is important.
3. We are proving a bound on the *expected* test error (or risk) as opposed to the usual high probability bound on the test error. Note that the test error is random because the hypothesis $A(S)$ is random.
4. Bounded expected risk implies agnostic PAC learnability, but we won't prove this in this lecture.
5. We use expected risk here because it's related to stability.

Example 8 (**revisit**)**.** Applying Tikhonov regularization to linear regression

$$A(X, \boldsymbol{y}) = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - y_i)^2$$

gives us ridge regression (from CSE 448/546):

$$A(X, \boldsymbol{y}) = \operatorname*{argmin}_{w \in \mathbb{R}^d}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - y_i)^2 + \lambda\|\boldsymbol{w}\|^2\right).$$

You can find a closed form solution by setting the gradient to zero and solving for $\boldsymbol{w}$.

Our goal is to show the following:

Theorem 2 (Ridge regression test risk bound)**.** Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times [-1, 1]$, where $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$. Let $\mathcal{H} = \{\boldsymbol{w} \in \mathbb{R}^d \mid \|w\| \leq B\}$. For any $\epsilon \in (0, 1)$, let $m \geq 150B^2/\epsilon^2$. Then, applying ridge regression with parameter $\lambda = \epsilon/(3B^2)$ satisfies

$$\mathbb{E}_{S \in D^m}[L_{\mathcal{D}}(A_{\mathrm{RLM}}(S))] \leq \min_{\boldsymbol{w} \in \mathcal{H}} L_{\mathcal{D}}(\boldsymbol{w}) + \epsilon.$$

8.1 Stability

By stability, we mean that "a small change in the input" does not "change the output much". In particular, on the input we consider two datasets that differ in a single example

$$\begin{aligned} S &= (z_1, z_2, \ldots, z_i, \ldots, z_n) \\ S^{(i)} &= (z_1, z_2, \ldots, z_i', \ldots, z_n) \end{aligned}$$

and we measure the effect as the loss on $z_i = (\boldsymbol{x}_i, y_i)$:

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i).$$

Intuitively, we expect

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \geq 0.$$

since $A(S^{(i)})$ does not get to see z_i while $A(S)$ does. If this difference is very large, then we might suspect the algorithm is overfitting because it is performing poorly on data it hasn't seen. On the other hand, if the estimator is stable, then this difference shouldn't be very big.

Now, let's define a precise notion of stability:

Definition 8.1 (on-average-replace-one-stability)**.** We say an algorithm A is on-average-replace-one-stable with rate $\epsilon(m)$ if

$$\mathbb{E}_{S,z',i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \epsilon(m)$$

for all $\mathcal{D}$, for some monotonically decreasing $\epsilon(\cdot)$.

This definition is justified by the following which shows that stable algorithms generalize:

Theorem 3. Let $S = (z_i, \ldots, z_n)$ be iid from $\mathcal{D}$ and $z' \sim \mathcal{D}$ another iid sample. Let $U[m]$ be the uniform distribution over $\{1, \ldots, m\}$. Then for any algorithm A,

$$\underbrace{\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]}_{\text{generalization gap}} = \underbrace{\mathbb{E}_{S,z',i\sim U[m]}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)]}_{\text{O.A.R.O.S.}}.$$

Proof.

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

and

$$\begin{aligned} \mathbb{E}_S[L_D(A(S))] &= \mathbb{E}_{S,z'}[\ell(A(S), z')] && \text{note that } S \text{ and } z' \text{ are independent} \\ &= \mathbb{E}_{S,z'}[\ell(A(S^{(i)}), z')] && \text{so we can swap them!} \end{aligned}$$

for all $i \in [m]$. □

8.2 Strong convexity

Assuming a convex loss function which is either Lipschitz or smooth, we show that RLM is stable because it is strongly convex.

Definition 8.2 (Strong convexity, line segment form)**.** A function is λ-strongly convex if for all $\boldsymbol{v}, \boldsymbol{w}$ and $\alpha \in (0, 1)$ we have

$$f(\alpha\boldsymbol{v} + (1-\alpha)\boldsymbol{w}) \leq \alpha f(\boldsymbol{v}) + (1-\alpha)f(\boldsymbol{w}) - \underbrace{\frac{\lambda}{2}\alpha(1-\alpha)\|\boldsymbol{v} - \boldsymbol{w}\|^2}_{\text{strong convexity term}}.$$

Remark 8.3. Every convex function is 0-strongly convex.

Remark 8.4. Personally, I find this definition a bit difficult to understand at first glance. If we name the RHS $g(\alpha)$ then note the following:

1. $g(0) = \boldsymbol{w}$.
2. $g(1) = \boldsymbol{v}$.
3. $\frac{\mathrm{d}^2 g}{\mathrm{d}\alpha^2} = \lambda\|\boldsymbol{v} - \boldsymbol{w}\|^2$ but this is in the α coordinate system. In the "C" coordinate system, the curvature is λ!

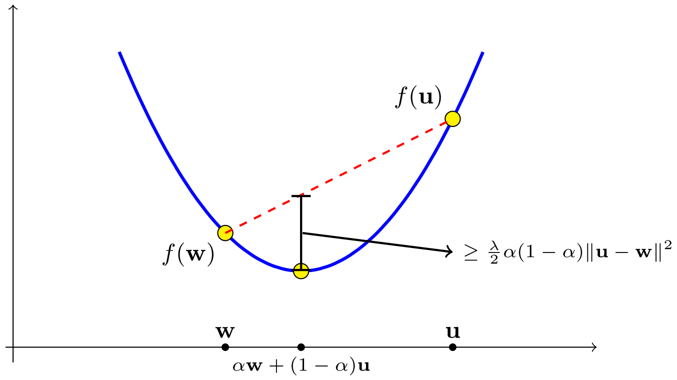
$f(\mathbf{u})$
$f(\mathbf{w})$
$\geq \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{u}-\mathbf{w}\|^2$
$\mathbf{w}$
$\mathbf{u}$
$\alpha\mathbf{w} + (1-\alpha)\mathbf{u}$

Figure 3: Graphical depiction of Proposition 8.2

For twice continuously differentiable functions, there is a nice test based on global minimum eigenvalue of the Hessian:

Proposition 8.5. Twice continuously differentiable f is λ-strongly convex iff $\nabla^2 f(\boldsymbol{w}) \succeq \lambda I$ for all $\boldsymbol{w}$.

Example 9. Here are some examples:

1. The function $f(\boldsymbol{w}) = \|\boldsymbol{w}\|^2$ is 2-strongly convex (and $f(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2$ is then clearly 2λ-strongly-convex).
2. Draw some function which is not strongly convex...
3. Give a sketch for why strong convexity + a nice loss gives stability of $A(S)$.

Proposition 9.1. If f is convex and g is λ-strongly convex then $f + g$ is λ-strongly convex.

These follow from the definition.

Now, recall that RLM was defined as

$$A(S) = \operatorname*{argmin}_{\boldsymbol{w}} \underbrace{\left(L_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2\right)}_{\text{denote by } f_S(\boldsymbol{w})}.$$

and from the above, we know that $f_S(\boldsymbol{w})$ is 2λ-strongly convex.

Proposition 9.2. If f is λ-strongly-convex, and $\boldsymbol{u}^*$ is the minimizer of f then for all $\boldsymbol{w}$,

$$f(\boldsymbol{w}) - f(\boldsymbol{u}^*) \geq \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{u}^*\|^2.$$

Proof. From the definition of strong convexity,

$$f(\boldsymbol{u}^*) \leq f(\alpha\boldsymbol{w} + (1-\alpha)\boldsymbol{u}^*) \leq \alpha f(\boldsymbol{w}) + (1-\alpha)f(\boldsymbol{u}^*) - \frac{\lambda}{2}\alpha(1-\alpha)\|\boldsymbol{w} - \boldsymbol{u}^*\|^2.$$

Collecting like terms and dividing by $\alpha > 0$

$$f(\boldsymbol{u}^*) \leq f(\boldsymbol{w}) - \frac{\lambda}{2}(1-\alpha)\|\boldsymbol{w} - \boldsymbol{u}^*\|^2.$$

Rearranging this and taking $\alpha \to 0^+$ gives the desired result. □

Let's try to bound $\|A(S^{(i)}) - A(S)\|$. We'll do this by bounding $f_S(A(S^{(i)})) - f_S(A(S))$ on both sides.

Lower bound: Note that $\boldsymbol{u}^* = A(S)$ in Proposition 9.2 so since $f_S(\boldsymbol{w})$ is 2λ-strongly convex, we have

$$f_S(A(S^{(i)})) - f_S(A(S)) \geq \lambda \|A(S^{(i)}) - A(S)\|^2.$$

Upper bound: We have

$$\begin{aligned} f_S(\underbrace{A(S^{(i)})}_{\hat{\boldsymbol{w}}^{(i)}}) - f_S(\underbrace{A(S)}_{\hat{\boldsymbol{w}}}) &= \underbrace{L_S(\hat{\boldsymbol{w}}^{(i)})}_{L_{S^{(i)}}(\hat{\boldsymbol{w}}^{(i)}) + \frac{\ell(\hat{\boldsymbol{w}}^{(i)}, z_i) - \ell(\hat{\boldsymbol{w}}^{(i)}, z')}{m}} + \lambda \|\hat{\boldsymbol{w}}^{(i)}\|^2 - \underbrace{L_S(\hat{\boldsymbol{w}})}_{L_{S^{(i)}}(\hat{\boldsymbol{w}}) + \frac{\ell(\hat{\boldsymbol{w}}, z_i) - \ell(\hat{\boldsymbol{w}}, z')}{m}} - \lambda \|\hat{\boldsymbol{w}}\|^2 \\ &= \underbrace{\big(\underbrace{L_{S^{(i)}}(\hat{\boldsymbol{w}}^{(i)}) + \lambda \|\hat{\boldsymbol{w}}^{(i)}\|^2}_{f_{S^{(i)}}(\hat{\boldsymbol{w}}^{(i)})} - \underbrace{L_{S^{(i)}}(\hat{\boldsymbol{w}}^{(i)}) + \lambda \|\hat{\boldsymbol{w}}^{(i)}\|^2}_{f_{S^{(i)}}(\hat{\boldsymbol{w}})}\big)}_{\leq 0 \text{ because } \hat{\boldsymbol{w}}^{(i)} \text{ minimizes } f_{S^{(i)}}} \\ &\quad + \frac{\ell(\hat{\boldsymbol{w}}^{(i)}, z_i) - \ell(\hat{\boldsymbol{w}}, z_i)}{m} - \frac{\ell(\hat{\boldsymbol{w}}^{(i)}, z') - \ell(\hat{\boldsymbol{w}}, z')}{m} \end{aligned}$$

Chaining the two inequalities together, we get

$$\lambda \|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} - \frac{\ell(A(S^{(i)}), z') - \ell(A(S), z')}{m}. \tag{1}$$

9.0.1 Lipschitz loss

If the loss function $\ell(\cdot, z)$ is ρ-Lipschitz for all z, then

$$\begin{aligned} &\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \rho \|A(S^{(i)}) - A(S)\|, \text{ and} \\ &\ell(A(S^{(i)}), z') - \ell(A(S), z') \leq \rho \|A(S^{(i)}) - A(S)\|. \end{aligned} \tag{2}$$

Plugging these both into (1), we get

$$\lambda \|A(S^{(i)}) - A(S)\|^2 \leq \frac{2\rho \|A(S^{(i)}) - A(S)\|}{m}.$$

Rearranging this yields

$$\|A(S^{(i)}) - A(S)\| \leq \frac{2\rho}{\lambda m}.$$

Plugging this into (2), we get

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{2\rho^2}{\lambda m}.$$

Since this holds for any S, z', i, we have the stability we desire:

Corollary 9.3. For convex and ρ-Lipschitz loss function, then RLM with $R(\boldsymbol{w}) = \lambda \|\boldsymbol{w}\|^2$ is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$.

Corollary 9.4. Applying Theorem 3, we obtain

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

Remark 9.5. The RHS is small when ρ is small (loss is less sensitive) and when λ is large (regularization is strong).

9.0.2 Smooth and Nonnegative Loss (revisit)

Proposition 9.6. Assume the loss function is β-smooth and nonnegative, then the RLM rule with the regularizer $\lambda\|\boldsymbol{w}\|^2$, where $\lambda \geq \frac{2\beta}{m}$, satisfies

$$\mathbb{E}_S[L_D(A(S)) - L_S(A(S))] \leq \frac{48\beta}{\lambda m}\,\mathbb{E}[L_S(A(S))].$$

9.1 Controlling the Fitting-Stability Tradeoff

Ok we bounded the generalization gap, but what we wanted is to bound the test loss. Let's expand

$$\begin{aligned}
\mathbb{E}_S[L_{\mathcal{D}}(A(S))] = \mathbb{E}_S[L_S(A(S))] + &\underbrace{\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))]}_{\text{bounded by stability!}} \\
&\leq \mathbb{E}_S[L_S(A(S)) + \lambda\|A(S)\|^2] + \frac{2\rho^2}{\lambda m} \\
&\leq \mathbb{E}_S[L_S(\boldsymbol{w}^*) + \lambda\|\boldsymbol{w}*\|^2] + \frac{2\rho^2}{\lambda m} && \text{for any } \boldsymbol{w}^* \\
&= L_{\mathcal{D}}(\boldsymbol{w}^*) + \underbrace{\lambda\|\boldsymbol{w}^*\|^2 + \frac{2\rho^2}{\lambda m}}_{\lambda \text{ trades off}}.
\end{aligned}$$

Corollary 9.7 (Convex-Lipschitz-bounded test risk bound)**.** If $\ell(\cdot, z)$ is convex and ρ-Lipschitz for all z and $\|\boldsymbol{w}^*\| \leq B$, then for $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$, RLM with regularization $R(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2$ satisfies

$$\mathbb{E}_S(L_{\mathcal{D}}(A(S))) \leq \min_{\boldsymbol{w}\in\mathcal{H}} L_{\mathcal{D}}(\boldsymbol{w}) + \rho B\sqrt{\frac{8}{m}}.$$

Remark 9.8. Solving for m, we see that if $m > 8\rho^2 B^2/\epsilon^2$ then for every distribution $\mathcal{D}$,

$$\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\boldsymbol{w}\in\mathcal{H}} L_{\mathcal{D}}(\boldsymbol{w}) + \epsilon.$$

This proves Theorem 1 after we note that with our bounds, logistic regression is 1-Lipschitz.

Now revisiting the convex-smooth-bounded case, we have

Corollary 9.9 (Convex-smooth-bounded test risk bound)**.** If $\ell(\cdot, z)$ is convex and β-smooth for all z, $\|\boldsymbol{w}^*\| \leq B$, and $\ell(\mathbf{0}, z) \leq 1$ for all z and for any $\epsilon \in (0, 1)$ we have $m \geq \frac{150\beta B^2}{\epsilon^2}$, then for $\lambda = \epsilon/(3B^2)$, RLM with regularization $R(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2$ satisfies

$$\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\boldsymbol{w}\in\mathcal{H}} L_{\mathcal{D}}(\boldsymbol{w}) + \epsilon.$$

This proves Theorem 2 after observing that in our setting, ridge regression is $\beta = 1$ smooth.