Lecture 5: Convex Learning and Regularization

Tianpei Xie

Feb. 10th., 2023

Contents

1 Convex Learning

1.1 Convexity, Lipschitzness and Smoothness

- **Definition** (***Convex Set***)
  A set C in a vector space is ***convex*** if for any two vectors u, v in C, ***the line segment*** between u and v is ***contained*** in C. That is, for any $\alpha \in [0,1]$ we have that $\alpha u + (1-\alpha)v \in C$.

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

  $$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v).$$

  A function f is ***convex*** if and only if its ***epigraph***

  $$\text{epi}(f) := \{(x,t) : t \geq f(x)\}$$

  is a ***convex set***.

- **Remark** (***Properties of Convex Function***)
  The important properties of convex function are as below:

  1. *Every* ***local minimum*** *of the function is also a* ***global minimum***. Formally, $f(u)$ is a ***local minimum*** of f at u if there exists some $r > 0$ such that for all $v \in B(u,r) := \{x : \|x - u\| < r\}$ we have $f(v) \geq f(u)$. It shows that if u is a local minima, then for every $v \in C$, $f(v) \geq f(u)$

  2. *Every w we can construct a* ***tangent*** *to f at w that* ***lies below*** *f* ***everywhere***. That is,

     $$f(u) \geq f(w) + \langle \nabla f(w),\, u - w \rangle, \quad \forall u \in C$$

  3. If f is second-order differentiable, then its Hessian matrix is ***positive semi-definite*** $\nabla^2 f \succeq 0$.

- **Definition** (***Lipschitz Function***)
  Let $C \subset \mathbb{R}^d$. A function $f : C \to \mathbb{R}^k$ is ***ρ-Lipschitz*** if there exists a constant $L > 0$ such that for every $u, v \in C$,

  $$\|f(u) - f(v)\| \leq \rho \|u - v\|.$$

  Intuitively, a *Lipschitz function cannot change too fast.*

- **Definition** (***Smoothness***).
  A ***differentiable*** function $f : \mathbb{R}^d \to \mathbb{R}$ is ***β-smooth*** if its ***gradient*** is β-*Lipschitz*; namely, for all u, v we have

  $$\|\nabla f(u) - \nabla f(v)\| \leq \beta \|u - v\|.$$

- **Remark** It is possible to show that if f is β-*smooth*, then for any u, v

  $$f(u) \leq f(v) + \nabla f(v)(u - v) + \frac{\beta}{2} \|u - v\|^2$$

Recall that *convexity* of f implies that

$$f(u) \geq f(v) + \nabla f(v)(u - v).$$

Therefore, when a function is ***both convex and smooth***, we have both *upper and lower bounds* on the difference between the function and its *first order approximation*.

- **Definition** (***Self-Boundedness***).
  A ***differentiable*** function $f : \mathbb{R}^d \to \mathbb{R}$ is ***self-bounded*** if the norm of its ***gradient*** is bounded by the function itself; namely, for all u we have

$$\|\nabla f(u)\|^2 \leq \gamma f(u).$$

  Note that a *β-smooth non-negative function* is *self-bounded*.

1.2 Convex Learning Problem

- **Definition** (***Convex Learning Problem***)
  A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called ***convex*** if *the* ***hypothesis class*** $\mathcal{H}$ *is a* ***convex set*** and ***the loss function***, $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, is a ***convex function*** *in its first argument* (where, for any z, the function $f : \mathcal{H} \to \mathbb{R}_+$ defined by $f(w) = \ell(w, z)$ is convex).

- **Lemma 1.1** *(**ERM of Convex Learning Problem is Convex**) [Shalev-Shwartz and Ben-David, 2014]*
  *If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the $ERM_{\mathcal{H}}$ problem, of minimizing **the empirical loss** over $\mathcal{H}$, is a **convex optimization problem** (that is, a problem of minimizing a convex function over a convex set)*

  Proof: Note that $ERM_{\mathcal{H}}$ find optimal hypothesis $h \in \mathcal{H}$ by minimizing the empirical loss

$$h = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

  By definition, $\ell(h, z_i)$ is a convex function in first argument and $\mathcal{H}$ is convex set, so the $ERM_{\mathcal{H}}$ is convex programming. ∎

- **Remark** *Note that **not all of convex learning problems** in $\mathbb{R}^n$ is **learnable*** even when n is low. That is, there exists some distribution $\mathcal{P}$ so that the generalization error is unbounded.

- **Definition** (***Convex-Lipschitz-Bounded Learning Problem***).
  A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called ***Convex-Lipschitz-Bounded***, with parameters ρ, B if the following holds:

  1. The hypothesis class $\mathcal{H}$ is a ***convex set*** and for all $w \in \mathcal{H}$ we have

$$\|w\|_{\mathcal{H}} \leq B.$$

  2. For all $z \in \mathcal{Z}$, the loss function, $\ell(\cdot, z)$, is a ***convex*** and ***ρ-Lipschitz function***:

$$|\ell(w, z) - \ell(v, z)| \leq \rho \|w - v\|_{\mathcal{H}}.$$

- **Definition** (***Convex-Smooth-Bounded Learning Problem***).
A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called ***Convex-Smooth-Bounded***, with parameters β, B if the following holds:

  1. The hypothesis class $\mathcal{H}$ is a ***convex set*** and for all $w \in \mathcal{H}$ we have
  $$\|w\|_{\mathcal{H}} \leq B.$$
  2. For all $z \in \mathcal{Z}$, the loss function, $\ell(\cdot, z)$, is a ***convex, non-negative*** and β***-smooth function***:
  $$|\nabla \ell(w, z) - \nabla \ell(v, z)| \leq \beta \|w - v\|_{\mathcal{H}}.$$
  Note that we also required that the loss function is ***non-negative***. This is needed to ensure that ***the loss function*** *is* ***self-bounded***.

1.3 Surrogate Loss Function

- **Remark** Convex problems can be learned effficiently. However, in many cases, the natural loss function is not convex and, in particular, implementing the ERM rule is hard.

  For instance, the 0-1 loss is *non-convex*:
  $$\ell^{0-1}(h, (X, Y)) := \mathbb{1}\{h(X) \neq Y\} = \mathbb{1}\{Yh(X) \leq 0\}.$$
  To circumvent the hardness result, one popular approach is to ***upper bound*** *the nonconvex loss function by a* ***convex surrogate loss function***. As its name indicates, the requirements from a convex surrogate loss are as follows:

  1. It should be ***convex***;
  2. It should ***upper bound*** the original loss.

- **Example** (***Hinge Loss***)
An example of convex surrogate function is ***the hinge loss***:
$$\ell^{hinge}(w, z) := \max\{0, 1 - y\langle w, x\rangle\}.$$
Clearly $\ell^{hinge}(w, z) := \max\{0, 1 - y\langle w, x\rangle\} \geq \mathbb{1}\{y\langle w, x\rangle \leq 0\} := \ell^{0-1}(w, z)$.

- **Remark** (***Decomposition of Loss***)
Given the surrogate loss $L^s_{\mathcal{P}}(h) = \mathbb{E}_{Z\sim\mathcal{P}}[\ell^s(h, Z)]$, we have decomposition of error
$$\underbrace{\left(L^s_{\mathcal{P}}(h_{\mathcal{D}}) - \min_{h\in\mathcal{H}} L^s_{\mathcal{P}}(h)\right)}_{\text{estimation error}} + \underbrace{\left(\min_{h\in\mathcal{H}} L^s_{\mathcal{P}}(h) - \min_{h\in\mathcal{H}} L^{0-1}_{\mathcal{P}}(h)\right)}_{\textbf{optimization error}} + \underbrace{\left(\min_{h\in\mathcal{H}} L^{0-1}_{\mathcal{P}}(h) - L^*\right)}_{\text{approximation error}}$$
where L^* is Bayes error and $h_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$ is the hypothesis returned by learning algorithm.

  The optimization error is a result of our inability to minimize the training loss with respect to the original loss. The size of this error depends on the specific distribution of the data and on the specific surrogate loss we are using.

2 Regularization

2.1 Regularized Loss Minimization

- **Definition** (*Regularized Loss Minimization (RLM)*)
  Regularized Loss Minimization (RLM) is a learning rule in which we jointly minimize ***the empirical risk*** *and a* ***regularization*** *function.* Formally, a ***regularization function*** is a mapping $R : \mathcal{H} \to \mathbb{R}_+$, and the regularized loss minimization rule outputs a hypothesis in

$$h \in \arg\min_{h \in \mathcal{H}} \left\{ L_{\mathcal{D}}(h) + R(h) \right\}. \tag{1}$$

- **Remark** (***Regularized Loss Minimization*** $\Rightarrow$ ***Structured Risk Minimization***)
  Regularized loss minimization shares similarities with ***minimum description length*** *algorithms* and ***structural risk minimization***. Intuitively, the "***complexity***" of hypotheses is measured by *the value of the regularization function*, and the algorithm balances between low empirical risk and "*simpler*," or "*less complex*," hypotheses.

- **Definition** (***Tikhonov Regularization***)
  Throughout this section we will focus on one of the most simple regularization functions: $R(w) = \lambda \|w\|_2^2$, where $\lambda > 0$ is a scalar and the norm is the ℓ_2 norm. The regualarized loss minimization problem becomes:

$$\mathcal{A}(\mathcal{D}) = \arg\min_{w \in C} \left\{ L_{\mathcal{D}}(w) + \lambda \|w\|_2^2 \right\}. \tag{2}$$

  This type of regularization function is often called ***Tikhonov regularization***.

- **Remark** Recall that in the previous section we introduced the notion of ***bounded hypothesis classes***. Therefore, we can define a sequence of hypothesis classes, $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \ldots$, where $\mathcal{H}_i = \{w : \|w\|_2 \leq i\}$. If the *sample complexity* of each $\mathcal{H}_i$ depends on i then *the RLM rule* is similar to *the SRM rule* for this sequence of nested classes.

2.2 Stable Rules Do Not Overfit

- **Remark** (***Stability***)
  Intuitively, a learning algorithm is ***stable*** if a small change of the input to the algorithm does not change the output of the algorithm much.

  Let $\mathcal{A}$ be a learning algorithm, let $\mathcal{D} = (z_1, \ldots, z_m)$ be a training set of m examples, and let $\mathcal{A}(\mathcal{D})$ denote the output of $\mathcal{A}$. Let $\widetilde{\mathcal{D}}^{(i)}$ be the training set obtained by replacing the i'th example of $\mathcal{D}$ with an i.i.d. copy $z_i' \neq z_i$, i.e.

$$\widetilde{\mathcal{D}}^{(i)} = \left(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m\right).$$

  The algorithm $\mathcal{A}$ is ***stable*** if for any $z_i \in \mathcal{Z}$,

$$0 \leq \ell\left(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}), z_i\right) \leq \epsilon_i \tag{3}$$

  That is, replacing any training sample will not cause significant increase of training loss.

- **Theorem 2.1** *(**Stability Implies Learnability**) [Shalev-Shwartz and Ben-David, 2014] Let $\mathcal{P}$ be a distribution. Let $\mathcal{D} = (z_1, \ldots, z_m)$ be an i.i.d. sequence of examples and let z' be another i.i.d. example. Then, for any learning algorithm,*

$$\mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D})) - L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right] = \mathbb{E}_{(\mathcal{D},z')}\left[\frac{1}{m}\sum_{i=1}^{m}\left\{\ell\left(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}), z_i\right)\right\}\right] \tag{4}$$

Proof: The expected generalization loss can be rewritten as

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] &:= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{Z\sim\mathcal{P}}\left[\ell(w_{\mathcal{D}}, Z)\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{Z'\sim\mathcal{P}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(w_{\mathcal{D}}, Z_i')\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{Z'\sim\mathcal{P}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(w_{\widetilde{\mathcal{D}}^{(i)}}, Z_i)\right]\right] \\
&:= \mathbb{E}_{\mathcal{D},Z'}\left[\frac{1}{m}\sum_{i=1}^{m}\ell\left(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), Z_i\right)\right]
\end{aligned}$$

The last equality holds since $w_{\mathcal{D}}, Z_i'$ are independent and Z_i and Z_i' are i.i.d. which means that exchanging Z_i with Z_i' will not change the expectation value. Moreover, the training loss can be written as

$$\mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right] := \mathbb{E}_{\mathcal{D}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell\left(\mathcal{A}(\mathcal{D}), Z_i\right)\right].$$

Thus proof the result. ∎

- **Definition** (***On-Average-Replace-One-Stable***). [Shalev-Shwartz and Ben-David, 2014] Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm $\mathcal{A}$ is ***on-average-replaceone-stable*** with rate $\epsilon(m)$ if for every distribution $\mathcal{P}$

$$\mathbb{E}_{(\mathcal{D},z')}\left[\frac{1}{m}\sum_{i=1}^{m}\left\{\ell\left(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}), z_i\right)\right\}\right] \leq \epsilon(m) \tag{5}$$

2.3 Tikhonov Regularization and Stability

- The main property of *the Tikhonov regularization* that we rely on is that it makes *the objective of RLM* ***strongly convex***

  Definition (***Strongly Convex Functions***).
  A function f is ***λ-strongly convex*** if for all w, u, and $\alpha \in (0, 1)$ we have

$$f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha)f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\left\|w - u\right\|_2^2.$$

- **Lemma 2.2** *(**RLM is 2λ-Strongly Convex**) [Shalev-Shwartz and Ben-David, 2014]*

  1. *The function $f(w) = \lambda\left\|w\right\|_2^2$ is **2λ-strongly convex**.*

2. *If f is λ-**strongly convex** and g **is convex**, then $f+g$ is λ-**strongly convex**.*

3. *If f is λ-**strongly convex** and u is a **minimizer** of f, then, for any w,*

$$|f(w) - f(u)| \geq \frac{\lambda}{2} \|w - u\|_2^2 .$$

Proof: We only prove the last part. By λ-strong-convexity of f

$$\frac{f\left(u + \alpha(w-u)\right) - f(u)}{\alpha} \leq (f(w) - f(u)) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2$$

Take limit as $\alpha \to 0$, the LHS becomes

$$\lim_{\alpha \to 0} \frac{f\left(u + \alpha(w-u)\right) - f(u)}{\alpha} = \nabla f(u)$$

Since u is the minimizer of f, $\nabla f(u) = 0$. The RHS becomes $(f(w) - f(u)) - \frac{\lambda}{2}\|w-u\|^2$. So the inequality becomes

$$(f(w) - f(u)) - \frac{\lambda}{2}\|w-u\|^2 \geq 0. \quad \blacksquare$$

- **Corollary 2.3** *(**RLM with Convex-Lipschitz is Stable**) [Shalev-Shwartz and Ben-David, 2014]*
*Assume that the loss function is **convex** and ρ-**Lipschitz**. Then, the RLM rule with the regularizer $\lambda\|w\|_2^2$ is **on-average-replace-one-stable** with rate $2\rho^2/(\lambda m)$. Then*

$$\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D})) - L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right] \leq \frac{2\rho^2}{\lambda m}. \tag{6}$$

Proof: Consider $\widetilde{\mathcal{D}}^{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)$ with z_i' is an i.i.d. copy of z_i. Let

$$\mathcal{A}(\mathcal{D}) := \arg\min_w \left\{ L_{\mathcal{D}}(w) + \lambda\|w\|_2^2 \right\}.$$

Denote $f_{\mathcal{D}}(w) = L_{\mathcal{D}}(w) + \lambda\|w\|_2^2$. Since $L_{\mathcal{D}}(w)$ is convex function and $\lambda\|w\|_2^2$ is 2λ-strongly convex, so $f_{\mathcal{D}}(w)$ is 2λ-strongly convex. Thus by *Lemma* above, for any w,

$$|f_{\mathcal{D}}(w) - f_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))| \geq \lambda\|w - \mathcal{A}(\mathcal{D})\|^2 . \tag{7}$$

On the other hand, for any v and u, and for all i, we have

$$\begin{aligned} f_{\mathcal{D}}(w) - f_{\mathcal{D}}(u) &= L_{\mathcal{D}}(w) + \lambda\|w\|_2^2 - L_{\mathcal{D}}(u) - \lambda\|u\|_2^2 \\ &= L_{\widetilde{\mathcal{D}}^{(i)}}(w) - L_{\widetilde{\mathcal{D}}^{(i)}}(u) + \lambda\left(\|w\|_2^2 - \|u\|_2^2\right) + \frac{\ell(w,z_i) - \ell(u,z_i)}{m} - \frac{\ell(w,z_i') - \ell(u,z_i')}{m} \end{aligned}$$

In particular, choosing $w = \mathcal{A}(\widetilde{\mathcal{D}}^{(i)})$, $u = \mathcal{A}(\mathcal{D})$, and using the fact that w minimizes $L_{\widetilde{\mathcal{D}}^{(i)}}(w) + \lambda\|w\|_2^2$, we obtain that

$$f_{\mathcal{D}}(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)})) - f_{\mathcal{D}}(\mathcal{A}(\mathcal{D})) \leq \frac{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i)}{m} - \frac{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i') - \ell(\mathcal{A}(\mathcal{D}), z_i')}{m} \tag{8}$$

Combining with (7), we have

$$\lambda \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\|^2 \leq \frac{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i)}{m} + \frac{\ell(\mathcal{A}(\mathcal{D}), z_i') - \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}{m} \tag{9}$$

Since ℓ is ρ-Lipschitz function, we have

$$\left| \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) \right| \leq \rho \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\| \tag{10}$$

Similarly the above inequality holds for z_i'. Substituting (10) into (9) we have

$$\begin{aligned} &\lambda \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\|^2 \leq 2\rho \frac{\left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\|}{m} \\ &\Rightarrow \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\| \leq \frac{2\rho}{\lambda m} \end{aligned}$$

Plugging back into (10), we have for all z_i,

$$\left| \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) \right| \leq \frac{2\rho^2}{\lambda m}.$$

By Theorem 2.1, we have the result. ∎

- **Corollary 2.4** *(**RLM with Convex-Smooth is Stable**) [Shalev-Shwartz and Ben-David, 2014]*
*Assume that the loss function is β-**smooth** and **nonnegative**. Then, the RLM rule with the regularizer $\lambda \|w\|_2^2$, where $\lambda \geq 2\beta/m$, satisfies*

$$\mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \left\{ \ell\left(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i \right) - \ell\left(\mathcal{A}(\mathcal{D}), z_i \right) \right\} \right] \leq \frac{32\beta}{\lambda m} \mathbb{E}\left[L_{\mathcal{D}}(\mathcal{A}(\mathcal{D})) \right]. \tag{11}$$

Note that if for all z we have $\ell(0, z) \leq B$, for some scalar $B > 0$, then for every $\mathcal{D}$,

$$L_{\mathcal{D}}(\mathcal{A}(\mathcal{D})) \leq L_{\mathcal{D}}(\mathcal{A}(\mathcal{D})) + \lambda \|\mathcal{A}(\mathcal{D})\|_2^2 \leq L_{\mathcal{D}}(0) + \lambda \|0\|_2^2 = L_{\mathcal{D}}(0) \leq B.$$

Proof: Define $f_{\mathcal{D}}(w) = L_{\mathcal{D}}(w) + \lambda \|w\|_2^2$. Following the same argument as previous proof, we have

$$\lambda \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\|^2 \leq \frac{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i)}{m} + \frac{\ell(\mathcal{A}(\mathcal{D}), z_i') - \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}{m} \tag{12}$$

We need to bound *the RHS*. Note that β-smooth and non-negative function are 2β-self-bounded

$$\|\nabla f(w)\|^2 \leq 2\beta f(w).$$

For $\lambda \geq 2\beta/m$, that is $\beta \leq \frac{m\lambda}{2}$. By smoothness assumption,

$$\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) \leq \left\langle \nabla \ell(\mathcal{A}(\mathcal{D}), z_i), \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\rangle + \frac{\beta}{2} \left\| \mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D}) \right\|^2 \tag{13}$$

By Cauchy-Schwartz inequality,

$$\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) \leq \|\nabla \ell(\mathcal{A}(\mathcal{D}), z_i)\| \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| + \frac{\beta}{2} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\|^2$$

By self-boundedness,

$$\|\nabla \ell(\mathcal{A}(\mathcal{D}), z_i)\|^2 \leq 2\beta \ell(\mathcal{A}(\mathcal{D}), z_i).$$

So

$$\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) \leq \sqrt{2\beta \ell(\mathcal{A}(\mathcal{D}), z_i)} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| + \frac{\beta}{2} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\|^2 \tag{14}$$

By a symmetric argument it holds that

$$\ell(\mathcal{A}(\mathcal{D}), z_i') - \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i') \leq \sqrt{2\beta \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| + \frac{\beta}{2} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\|^2$$

Plugging these inequalities into Equation (12) and rearranging terms we obtain that

$$\left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| \leq \frac{\sqrt{2\beta}}{m\lambda - \beta} \left[\sqrt{\ell(\mathcal{A}(\mathcal{D}), z_i)} + \sqrt{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}\right]$$

Since $\beta \leq \frac{m\lambda}{2}$,

$$\left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| \leq \frac{\sqrt{8\beta}}{m\lambda} \left[\sqrt{\ell(\mathcal{A}(\mathcal{D}), z_i)} + \sqrt{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}\right] \tag{15}$$

Combing (15) with (14), we have

$$\begin{aligned}
\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i) - \ell(\mathcal{A}(\mathcal{D}), z_i) &\leq \sqrt{2\beta \ell(\mathcal{A}(\mathcal{D}), z_i)} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\| \\
&\quad + \frac{\beta}{2} \left\|\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}) - \mathcal{A}(\mathcal{D})\right\|^2 \\
&\leq \left(\frac{4\beta}{m\lambda} + \frac{8\beta^2}{(m\lambda)^2}\right) \left(\sqrt{\ell(\mathcal{A}(\mathcal{D}), z_i)} + \sqrt{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}\right)^2 \\
&\leq \frac{8\beta}{m\lambda} \left(\sqrt{\ell(\mathcal{A}(\mathcal{D}), z_i)} + \sqrt{\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')}\right)^2 \\
&\leq \frac{16\beta}{m\lambda} \left(\ell(\mathcal{A}(\mathcal{D}), z_i) + \ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')\right)
\end{aligned}$$

where in the last inequality we use $(a+b)^2 \leq 2(a^2+b^2)$. Taking expectation on both sides and noting that $\mathbb{E}\left[\ell(\mathcal{A}(\mathcal{D}), z_i)\right] = \mathbb{E}\left[\ell(\mathcal{A}(\widetilde{\mathcal{D}}^{(i)}), z_i')\right] = \mathbb{E}\left[L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right]$, we have the result. ∎

2.4 PAC Learnability for Convex Learning Problem

- Note that

$$\mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] = \mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right] + \mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D})) - L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right].$$

  So the stability result will lead to learnability result.

Corollary 2.5 ***(Oracle Inequality of RLM for Convex-Lipschitz-Bounded Problem)*** *[Shalev-Shwartz and Ben-David, 2014]*
Assume that the loss function is ***convex*** *and* ***ρ-Lipschitz****. Then, the* ***RLM rule*** *with the regularizer $\lambda \|h\|^2$ satisfies*

$$\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] \leq L_{\mathcal{D}}(h^*) + \lambda \|h^*\|^2 + \frac{2\rho^2}{\lambda m}. \tag{16}$$

for some fixed arbitrary rule h^*. This inequality is called ***the oracle inequality****, since h^* is thought as a hypothesis with low risk.*

- We can also easily derive a *PAC-like guarantee* for *convex-Lipschitz-bounded learning problems*:

  Corollary 2.6 ***(PAC Learnability of Convex-Lipschitz-Bounded Problem)*** *[Shalev-Shwartz and Ben-David, 2014]*
  Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be a ***convex-Lipschitz-bounded*** *learning problem with parameters ρ, B. For any training set size m, let*

$$\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}.$$

  Then, the ***RLM rule*** *with the regularizer $\lambda \|h\|^2$ satisfies*

$$\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{8B^2\rho^2}{m}} \tag{17}$$

  In particular, for every $\epsilon > 0$, if $m \geq \frac{8\rho^2 B^2}{\epsilon^2}$ then for every distribution $\mathcal{P}$, $\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

- **Corollary 2.7** ***(Oracle Inequality of RLM for Convex-Smooth-Bounded Problem)*** *[Shalev-Shwartz and Ben-David, 2014]*
  Assume that the loss function is ***convex, β-smooth, and nonnegative****. Then, the* ***RLM rule*** *with the regularizer $\lambda \|h\|^2$, for $\lambda \geq \frac{2\beta}{m}$, satisfies the following for all h^**

$$\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] \leq \left(1 + \frac{32\beta}{\lambda m}\right) \mathbb{E}\left[L_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))\right] \leq \left(1 + \frac{32\beta}{\lambda m}\right)\left(L_{\mathcal{D}}(h^*) + \lambda \|h^*\|_2^2\right). \tag{18}$$

- **Corollary 2.8** ***(PAC Learnability of Convex-Smooth-Bounded Problem)*** *[Shalev-Shwartz and Ben-David, 2014]*
  Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be a ***convex-smooth-bounded*** *learning problem with parameters β, B. Assume in addition that $\ell(0, z) \leq 1$ for all $z \in \mathcal{Z}$. For any $\epsilon \in (0, 1)$ let*

$$m \geq \frac{150\beta B^2}{\epsilon^2}, \quad \textit{and} \quad \lambda = \frac{\epsilon}{3B^2},$$

  then for every distribution $\mathcal{P}$, $\mathbb{E}\left[L_{\mathcal{P}}(\mathcal{A}(\mathcal{D}))\right] \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

References

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.