## Remider:

$$\hat{R}_S[h] = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i); y_i) \leftarrow \text{empirical risk}$$

$$R[h] = E_{z\sim D}\,\ell(h(x); y) \leftarrow \text{true risk}$$

## Last time:

Empirical risk concentration.

- Main message: Empirical risk is within $\varepsilon$ from true if

$$\#\text{samples} \geq O\left(\frac{\#\text{params}}{\varepsilon^2}\right)$$

However, There are very sophisticated techniques to extend this beyond finite classes. eg.
- VC-dimension
- Rademacher complexity.

In future lectures we will see concentration bounds that are "algorithm-specific"

## Lecture 3:

- From statistical bounds to optimization
- Computational aspects of the ERM
- Examples of loss fns

---

What do concentration bounds tell us?

- Lets assume that

$$\hat{R}_S[h] \leq R[h] + \varepsilon, \quad \forall h \in \mathcal{H}$$

with prob. $1-\delta$

But we care about a "special" $h$:

- $\hat{h}^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}[h] \leftarrow$ ERM

and its true performance

$$R[\hat{h}^*] = E_z \, \ell(\hat{h}^*(x); y)$$

Then, if we have concentration $\forall h \in \mathcal{H}$

$$\Rightarrow \quad R[\hat{h}^*] = R[\hat{h}^*] + (\hat{R}[\hat{h}^*] - R[\hat{h}^*])$$

$$= \hat{R}[\hat{h}^*] + (R[\hat{h}^*] - \hat{R}[\hat{h}^*])$$

$$\leq \hat{R}[\hat{h}^*] + \varepsilon \quad \text{w.p. } 1-\delta$$

If the ER concentrates then

$$R[\hat{h}^*] \le \hat{R}[\hat{h}^*] + \varepsilon$$

But also we can relate $R[\hat{h}^*]$ to the best predictor in $\mathcal{H}$

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R[h]$$

We have that

$$\begin{aligned} R[\hat{h}^*] &\le \hat{R}[\hat{h}^*] + \varepsilon \\ &\le \hat{R}[h^*] + \varepsilon \\ &\le R[h^*] + \underbrace{\hat{R}[h^*] - R[h^*]}_{\varepsilon} + \varepsilon \\ &\le R[h^*] + 2\varepsilon \end{aligned}$$

Hence, we can argue about the best possible predictor via the performance of the ERM, assuming concentration

The above are a brief preview of "why ERM is a good idea".

But what does it look like?

Examples:

- Regression:
  - linear: $\min_{w} \|Xw - y\|^2$ $\left[\begin{array}{l} +\lambda\|w\|_2^2 \text{ ridge} \\ +\lambda\|w\|_1 \text{ lasso} \end{array}\right]$

    $= \min_{w} \frac{1}{n} \sum (x_i^T w - y_i)^2$
  - nonlinear (eg. nn)

    $\min \frac{1}{n} \sum (y_i - h(x_i; w))^2$

    $h(x_i; W) = \sigma(W_L\, \sigma(W_{L-1} \sigma( \dots \sigma(W_1 x))$

- Classification:
  - Binary: $\min \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i x_i^T w})$

    $\min \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i x_i^T w\}$

- Multiclass: (for one sample)

$$-\sum_{c=1}^{M} y_{i,c} \log([h(w;x_i)]_c)$$

$\vdots$

## Back to optimization:

We want to solve

$$\min_{w} \frac{1}{n}\sum_{i=1}^{n} l_i(w) + \lambda \cdot R(w)$$

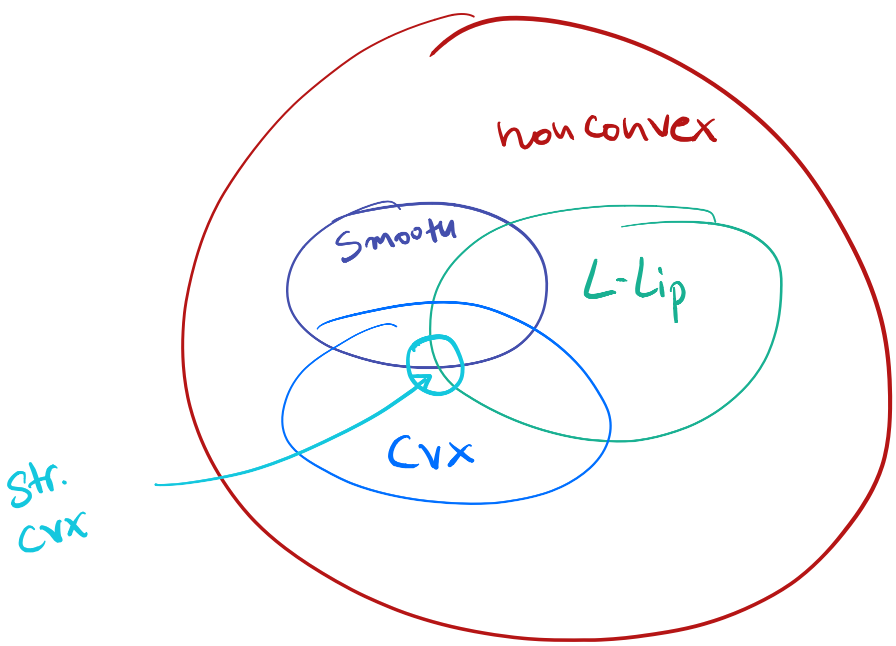
- Q: When can we solve it?
- Q: How fast?

Remark: The more you know about the structure of the problem the more we can say about "solvability" and "scalability".

Informal Theorem: In the general case, ERM is NP-Hard.

Let's put some structure

Families of functions

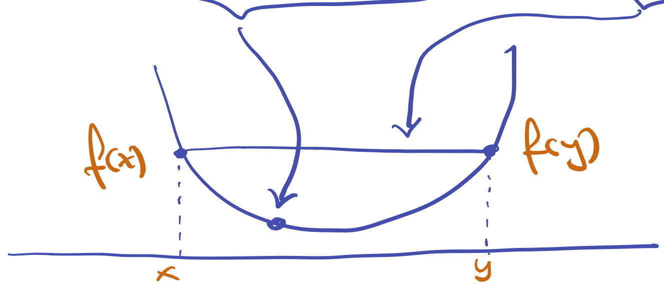



• **Convex:**

$$f(a\cdot\vec{x}+(1-a)\vec{y}) \le a\cdot f(\vec{x})+(1-a)f(\vec{y})$$

also

$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle$




Convexity makes our lives easy, eg

$\min f(w)$ solvable in poly-time

• Important property of cvx functions:

Every local min $\equiv$ global min

L- Lipschitz: ("fns that don't change fast")

$\forall \; x,y \quad |f(x)-f(y)| \leq L\|x-y\|$

β-smooth: ("fns with grads that don't change fast")

$\forall \; x,y \quad \|\nabla f(x)-\nabla f(y)\| \leq \beta\|x-y\|$

strongly convex: ("the best kind of fns")

$$f(x)-f(y) \leq \langle \nabla f(x), x-y \rangle - \frac{\alpha}{2}\|x-y\|^2$$

## Examples:

- **convex:**
  - $\|x\|^2$, $\|x\|$, $\log(1+\exp(x))$, $\max\{0, 1-x\}$
  - if $g(\cdot)$ is cvx, then $g(x^T w+b)$ is
    eg: $\log(1+\exp(-y\langle w,x\rangle))$
    $(w^T x - b)^2 \ldots$
  - $\max_i f_i(x)$ (if $f_i$ are cvx)
  - $\sum_i w_i f_i(x)$ (if $f_i$ are cvx)
  - $\ldots$
- **L-Lip:**
  - $|x|$ is 1-Lip.
    $f(x) = \log(1+\exp(x))$ is 1-Lip
  - $x^2$ is not Lipschitz
    unless $|x| \le p$ when it is $p$-Lip
  - $f(w) = x^T w + b$ is $\|x\|$-Lip.
  - $f(x) = g_1(g_2(x))$. If $g_1$ is $L_1$-Lip.
    $g_2$ is $L_2$-Lip
    $\Rightarrow f$ is $L_1 \cdot L_2$-Lip.

- $g(x^T w + b) \Rightarrow \|x\| \cdot L_g$ - Lip.
- If $\|\nabla f(w)\| \leq L \Rightarrow f$ is $L$-Lip.

- Smooth:
  - $|x|^2$ is 2-smooth
  - $\log(1+e^x)$ is $\frac{1}{4}$-smooth
  - If $g$ is $\beta$-smooth
    $f(w) = g(w^T x + b)$ is $\beta \|x\|^2$-smooth
  - $f(w) = \log(1 + e^{-y\langle w, x\rangle})$ $\frac{\|x\|^2}{4}$-smooth

- Strongly Convex:
  - $f$ is $\lambda$-str convex if
    $f(w) - \frac{\lambda}{2}\|w\|^2$ is convex
  - eg. $\sum_{i=1}^{n} \log(1 + e^{-y\langle w, x\rangle}) + \frac{\lambda}{2}\|w\|^2$

    $\vdots$

Next time:

- Why is convexity useful?
- How to exploit it algorithmically?
- Gradient Methods