$$\hat{f}_n = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{\hat{P}_n} \ell(f, \xi) \qquad \hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}, \quad \text{e.g. } \xi_i = [x_i, y_i]$$

$$\hat{L}_f$$

Hypothesis class

$$\Big(\int \qquad d\hat{P}_n(\xi) = \frac{1}{n}\sum_{i=1}^{n} \xi_i \qquad \xi_i \overset{iid}{\sim} P_0$$

$$E_{P_0} \ell(\hat{f}_n, \xi)$$

Excess risk vs.

$$E_{P_0} \ell(\hat{f}_n, \xi) - \inf_{f \in \mathcal{H}} E_{P_0} \ell(f, \xi)$$

$$\Big(\text{CLT} \quad \frac{1}{n}\sum_{i=1}^{n} z_i \longrightarrow \Big) \qquad L_{\hat{f}} - L_f$$

$$L_{\hat f} - L_{f^*} \ ? \qquad \inf_{f \in \mathcal{H}} \mathbb{E}_{P_0} \ell(f, z)$$

$$= \underbrace{L_{\hat f} - \hat L_{\hat f}}_{①} + \underbrace{\hat L_{\hat f} - \hat L_{f^*}}_{② \ \le 0} + \underbrace{\hat L_{f^*} - L_{f^*}}_{③}$$

$$\sup_{f \in \mathcal{H}} \left\{ \mathbb{E}_{P_0} \ell_f - \frac{1}{n} \sum_{i=1}^{n} \ell_f \right\}$$

Concentration

Linear model

$$f(x) = \sum_{i=1}^{d} w_i x_i + b$$

$\xi_i \sim P_0$

$= [x_i, y_i]$

- Least-Squares Regression

$$\boxed{\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i) - y_i)^2}$$

$\left\|\begin{matrix} w \\ b \end{matrix}\right\| \le B^2$

$\sum_{i=1}^{d} |w_i|^2$

$$\boxed{B \mid \|w\|_2^2 + b^2 \le B^2}$$

f^{-1}

Generalization:

$$\boxed{L_f^{\hat{}} - L_f \le O\left(\frac{1}{\sqrt{n}}\right) + \frac{B \cdot C}{\sqrt{n}} \quad (\text{w.l.p})}$$

$\cancel{O\left(\frac{1}{n}\right)}$

$(XX^T)^{-1} X Y^T$

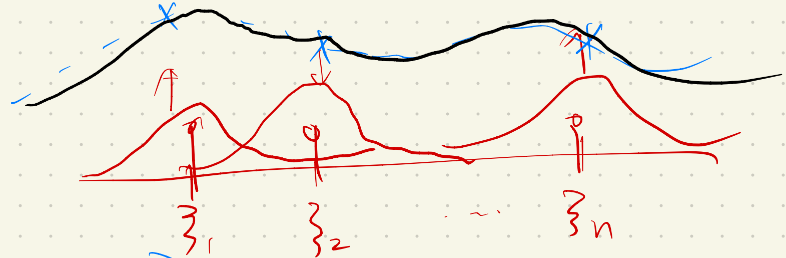
$O(n^2) - O(n^3)$

$O(n^{1.5})$

- Logistic regression

Kernel Functions ~~(RKHS)~~

$$k(x, x') = e^{-\frac{|x-x'|^2}{2}}$$



$$\mathcal{H} = \text{comp}\left\{\text{span}\left\{ k(x, \cdot) \;\middle|\; x \in \mathcal{X} \, (\mathbb{R}^d) \right\}\right\}$$

$$\sum_{i=1}^{n_f} w_i \phi(x)$$

$$f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \quad , \quad g = \sum_{i=1}^{m} \beta_i k(x_i', \cdot)$$

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{j}^{m} \sum_{i}^{n} \alpha_i \beta_j k(x_i, x_j')$$

$$\|f\|_{\mathcal{H}} = \sqrt{\alpha^T K \alpha}$$

$\forall f \in \mathcal{H}$.

$k(x_i, x)$

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$$

$\in \mathcal{H}$

$k(x, \cdot)$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\hat{P}_n} \ell(f(x), y)$$

$$B \mid \|f\|_{\mathcal{H}}^2 \leq B^2$$

$$L_{\hat{f}_n} - L_f \leq O\left(\frac{1}{\sqrt{n}}\right) + \frac{B \cdot C}{\sqrt{n}}$$

$$f = \sum_{i=1}^{n} \underline{\alpha_i} k(x_i, \cdot)$$

$$\sqrt{\alpha^T K \alpha} =$$

$$\alpha^T Q \Lambda \underline{Q' \alpha} = u \Lambda u' \quad u u' = \alpha \alpha'$$

$$\frac{\lambda_i \propto r^i}{\propto i^{-t}} \quad \sum |\lambda_i|_0 < \infty$$

Thm (Generalization of 2-layer NN)

$$f(x) = W_2^T \underline{b}(\underline{W_1 x + b_1}) + b_2$$

$$\boxed{L_{\hat{f}_n} - L_f \leq O\left(\frac{1}{\sqrt{n}}\right) + \frac{B_1 B_2 \sqrt{m}}{\sqrt{n}}}$$

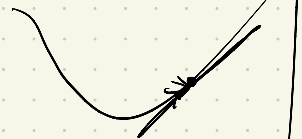
hidden units

$$\|W_i\|^2 + \|b_i\|^2 \leq B_i^2$$

obj.

$$\min_{x \in \mathcal{X}} f(x)$$

$(\mathbb{R}^d)$

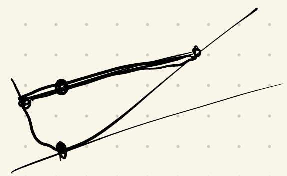


Def. (Convexity)

① f is convex iff

$\forall \lambda \in [0,1]$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y), \quad \forall x, y \in \mathcal{X}$$



② Sps f is diff.

$$f(x) \geq f(y) + \nabla f(y)(x-y)$$

(local optimum is global)

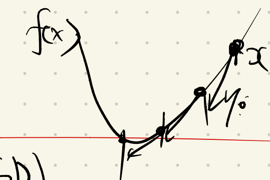
• LSQ ✓

• K(R)R ✓

$$\text{obj} + \lambda \|f\|_{\mathcal{H}}^2$$

$$\alpha^T K \alpha$$

• 2-L NN ✗



Alg Gradient descent method (GD)

- given x_0.
- at iteration $k = 0, \cdots, t-1$

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

Thm. Sps f is L-lip. R is such that $\|x_0 - x\| \leq R$, $\forall x$

Sps $\gamma_k = \frac{R}{L} \cdot \frac{1}{\sqrt{t}}$ then $f\left(\frac{1}{t}\sum_{k=1}^{t} x_k\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}$

$\forall x, y \in \mathcal{X}$

$|f(x) - f(y)| \le L\,|x-y|$

$|\nabla f(x) - \nabla f(y)| \le \beta\,|x-y|$ $\quad(\nabla^2 f \preceq \beta I)$

then f is L-Lipschitz

$\|\nabla f\| \le L$

$\sup_x \|\nabla f(x)\| = L$

Def. (α-str.) Convex

f is α-Strongly Convex if

$$f(x) - f(y) \le \nabla f(x)^T (x-y) - \frac{\alpha}{2}\|x-y\|^2$$

$$f(x) + \nabla f(x)(y-x) + \frac{\alpha}{2}\|y-x\|^2 \le f(y)$$

(α-convex)

$+\frac{\beta}{2}\|y-x\|^2$

Def. (β-Smooth)

f is β-Smooth if ∇f is $\boxed{\beta\text{-Lipschitz}}$.

$\bar{x} \in [x, y]$

$$\le f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)\nabla^2 f(\bar{x})(y-x)$$

Oracle complexity

- $\underline{f(x_k)}$
- $\underline{\nabla f(x_k)}$ $(\partial f(x_k))$

Elementary ops count (flops)

Thm (GD w/ str. cvx, smooth f)

Sps f is α-str. cvx, β-smooth,

then GD w/ $\eta_k = \frac{1}{\alpha+\beta}$ satisfies

$$f(x_k) - f(x^*) \le \frac{\beta}{2} e^{-\frac{4k}{\frac{\beta}{\alpha}+1}} \|x_1 - x^*\|^2 \quad (\le R)$$

$$\text{argmin} \left\{ f^k + \nabla f^k (x - x^k) + \frac{1}{2\eta} \|x - x^k\|^2 \right\}$$