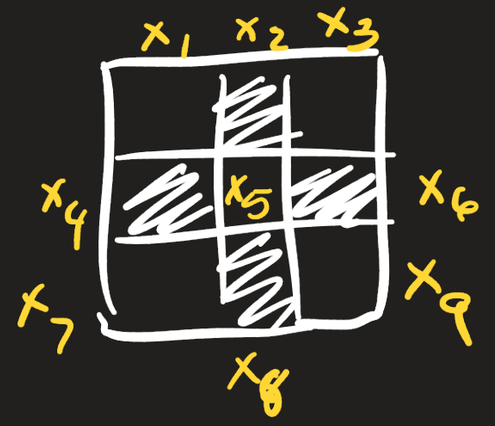
Lecture II Machine Learning & Optimization

- Kernel Learning
- Strong convexity and convex condition numbers
- Polyak-Łojasiewicz (PL) inequality & linear convergence of GD
- Convergence Rates for different classes of convex functions

Kernel Learning



zeros look like $[0,1,0,1,0,1,0,1,0]$

discriminative feature:

$$x_2 x_4 x_6 x_8$$

in general we expect polynomial combinations of the features are more useful for classification than linear combinations

Problem: there are $O(d^q)$ monomials in d variables of up to degree q

$$x \longmapsto \phi(x) = [1 \; x_1 \; x_2 \; \cdots \; x_d \; x_1^2 \; x_1 x_2 \; x_2^2 \; \cdots \; x_d^2 \; \cdots \; x_d^q] \in \mathbb{R}^{d^q}$$

fit a classifier for 0 vs 1

$$\min_{w \in \mathbb{R}^{d^q}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i w^T \phi(x_i)}\right) + \lambda \|w\|_2^2$$

Recall Fermat's condition for ERM w/ ℓ_2-regularization

$$\omega^* = \operatorname*{argmin}_{\omega \in \mathbb{R}^D} \frac{1}{n} \sum_{i=1}^{n} \ell(\omega^T \phi(x_i), y_i) + \lambda \|\omega\|_2^2$$

where $\phi : x \mapsto \phi(x)$, $\phi : \mathbb{R}^d \to \mathbb{R}^D$

$$0 \in \frac{1}{n} \sum_{i=1}^{n} \partial \ell(\omega^{*T} \phi(x_i), y_i) + 2\lambda \omega^*$$

i.e. there are $z_i \in \partial \ell(\omega^{*T} \phi(x_i), y_i)$ s.t.

$$0 = \frac{1}{n} \sum_{i=1}^{n} z_i \phi(x_i) + 2\lambda \omega^*$$

$$\Rightarrow \omega^* \in \operatorname{span}\{\phi(x_i)\}_{i=1}^{n}$$

This implies that if we solve

$$\alpha^* = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell\left(\left[\sum_{j=1}^{n} \alpha_j \phi(x_j)\right]^T \phi(x_i), y_i\right) + \lambda \left\| \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\|_2^2$$

notice

$$\left\langle \sum_{j=1}^{n} \alpha_j \phi(x_j), \phi(x_i) \right\rangle = \sum_{j=1}^{n} \alpha_j \underbrace{\langle \phi(x_j), \phi(x_i) \rangle}_{K(x_i, x_j)}$$

and

$$\left\| \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\|_2^2 = \left\langle \sum_{j=1}^{n} \alpha_j \phi(x_j), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_j \alpha_i \, K(x_i, x_j)$$

Define the <u>Kernel Matrix</u> $K \in \mathbb{R}^{n \times n}$

$$K = \left[k(x_i, x_j) \right]_{\substack{i=1,\dots,n \\ j=1,\dots,n}}$$

Notice

$$\sum_{j=1}^{n} k(x_i, x_j)\alpha_j = \sum_{j=1}^{n} K_{ij}\alpha_j = (K\alpha)_i$$

and $$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} = \alpha^T K \alpha$$

so we can find $\omega^* = \sum_{i=1}^{n} \alpha_i^* \phi(x_i)$ where

$$\alpha^* = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell((K\alpha)_i, y_i) + \lambda \alpha^T K \alpha$$

Two issues

1) evaluating K could require explicitly computing

$$K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle \quad \text{for all } i, j$$

However there are standard choices of ϕ that work well for which $K(x_i, x_j)$ can be computed in $O(d)$

$$K(x,y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}} \qquad \text{RBF - kernel}$$

$$K(x,y) = e^{-\frac{\|x-y\|_1}{B}} \qquad \text{Laplace - kernel}$$

$$K_{poly}(x,y) = (x^T y + c)^p \qquad \text{degree-p polynomial kernel}$$

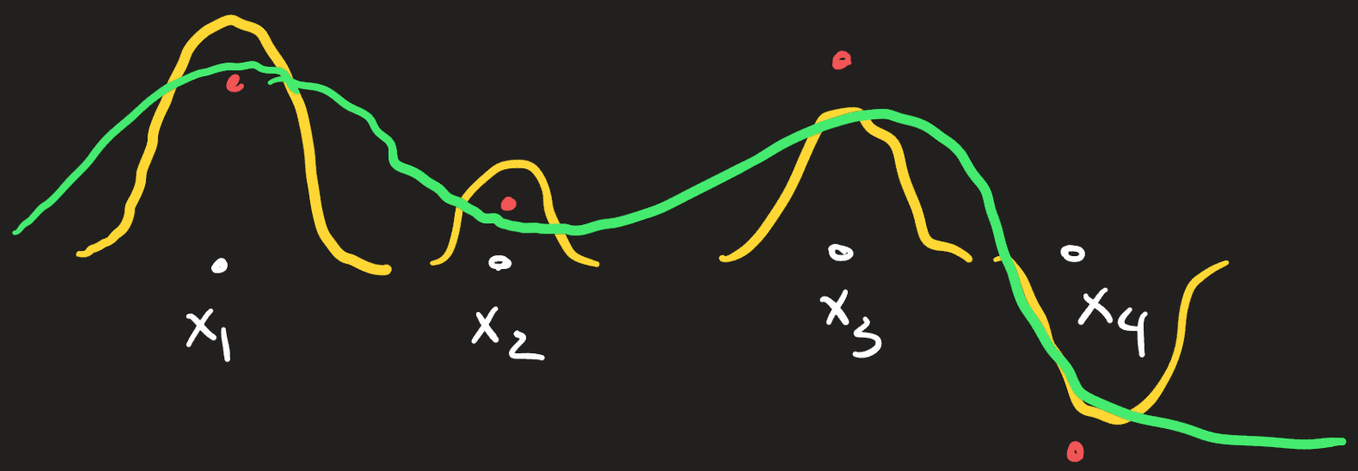
2) We still have to compute $w^* = \sum_{i=1}^{n} \alpha_i^* \phi(x_i)$

Solution: we only care about predictions!

$$\langle \phi(x), w^* \rangle = \sum_{i=1}^{n} \alpha_i^* \langle \phi(x), \phi(x_i) \rangle$$

$$= \sum_{i=1}^{n} \alpha_i^* K(x, x_i)$$

So we can predict using the kernel map

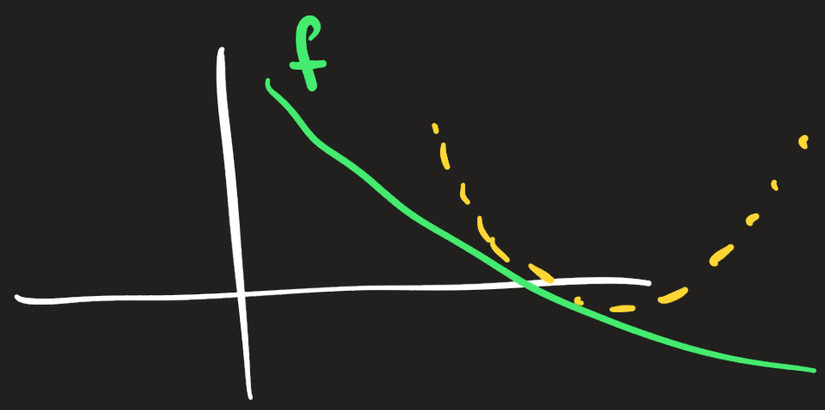


Recap: β-smoothness

A function f is β-smooth if $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x-y\|_2$

Consequence:

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \beta \frac{\|x-y\|_2^2}{2}$$



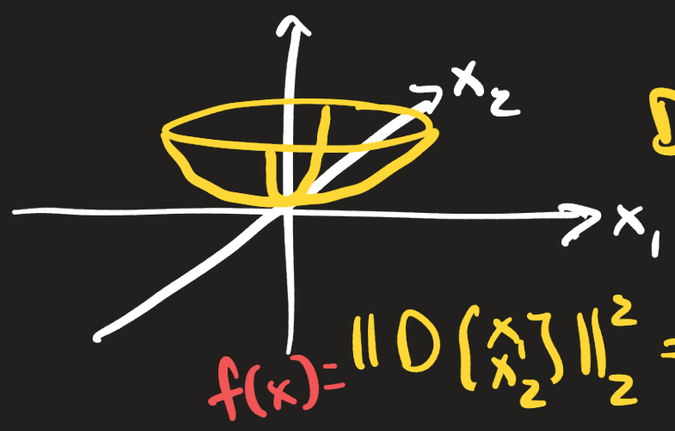
Consequence: (Descent Lemma)

$$x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t) \Rightarrow f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

α – strong convexity

We say a function is α-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|x-y\|_2^2 \qquad \forall x,y \in \text{dom}(f)$$



$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \quad \text{where } \lambda_1 \ll \lambda_2$$

$$\|Dy\|_2^2 = \|Dx\|_2^2 + \langle 2D^2x, y-x \rangle + \frac{1}{2}\langle y-x, 2D^2(y-x) \rangle$$

$$= \|Dx\|_2^2 + 2\langle D^2x, y-x \rangle + \|D(y-x)\|_2^2$$

$$= \|Dx\|_2^2 + 2\langle D^2x, y-x \rangle + \lambda_1(y_1-x_1)^2 + \lambda_2(y_2-x_2)^2$$

$$\geq \|Dx\|_2^2 + 2\langle D^2x, y-x \rangle + \frac{2\lambda_1}{2}\|y-x\|_2^2$$

so f is $2\lambda_1$ – strongly convex

Characterization

A twice-differentiable convex function f is

- α-strongly convex on its domain iff

$$\nabla^2 f(x) \succcurlyeq \alpha I \qquad \forall x \in \mathrm{dom}(f)$$

- β-smooth on its domain iff

$$\nabla^2 f(x) \preccurlyeq \beta I$$

Note that $\beta \geq \alpha$. We call $\kappa_f = \frac{\beta}{\alpha} \geq 1$ the convex condition number of f.

GD converges very fast for β-smooth, α-strongly convex functions

Claim: $f(x_T) - f(x_*) \leq e^{-T/\kappa_f} (f(x_0) - f(x^*))$

if we use GD w/ stepsize $\frac{1}{\beta}$

PL-inequality Strongly convex functions satisfy the PL-inequality:

$$\frac{1}{2\alpha} \|\nabla f(x)\|_2^2 \geq f(x) - f(x^*)$$

Pf

for all y,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x-y\|_2^2$$

$$\geq \min_z \; f(x) + \langle \nabla f(x), z - x \rangle + \frac{\alpha}{2}\|x-z\|_2^2$$

The RHS is a convex quadratic, so can be minimized by setting its gradient to zero

$$\nabla f(x) + \alpha(z-x) = 0$$

$$\Rightarrow \; z = x - \frac{1}{\alpha}\nabla f(x)$$

and gives

$$f(y) \geq f(x) + \langle \nabla f(x), -\frac{1}{\alpha}\nabla f(x) \rangle + \frac{\alpha}{2}\|\frac{1}{\alpha}\nabla f(x)\|_2^2$$

$$= f(x) - \frac{1}{\alpha}\|\nabla f(x)\|_2^2 + \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

That is, for all y,

$$f(y) \geq f(x) - \frac{1}{2\alpha} \|\nabla f(x)\|_2^2$$

Take $y = x^*$ and rearrange to see

$$\|\nabla f(x)\|_2^2 \geq 2\alpha \left[f(x) - f(x^*) \right]$$

Pf linear convergence of GD

Consider

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

by the descent lemma, so

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

and by the PL-inequality,

$$\|\nabla f(x_t)\|_2^2 \geq 2\alpha\left(f(x_t) - f(x^*)\right)$$

so

$$f(x_{t+1}) - f(x_t) \leq -\frac{\alpha}{\beta}\left(f(x_t) - f(x^*)\right)$$

adding $f(x_t) - f(x^*)$ to both sides,

$$f(x_{t+1}) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)\left[f(x_t) - f(x^*)\right]$$

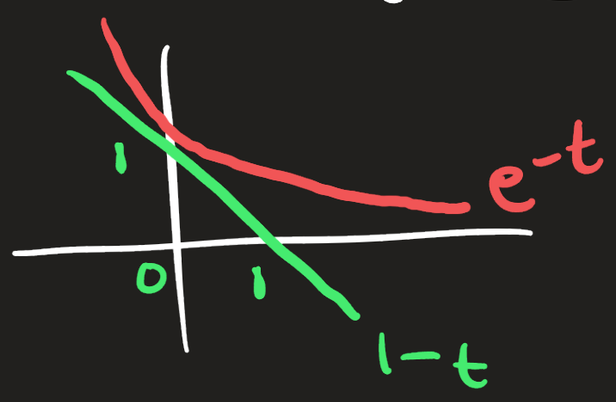
by induction

$$f(x_T) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^T \left(f(x_0) - f(x^*)\right)$$

$$= \left(1 - \frac{1}{\kappa_f}\right)^T \left(f(x_0) - f(x^*)\right)$$

$$\leq e^{-T/\kappa_f} \left(f(x_0) - f(x^*)\right)$$

fact $1 - t \leq e^{-t}$



∎

Convergence rates for different classes of cvx functions using GD

properties	cvx	β-smooth	α-strongly cvx	$\kappa_f = \beta/\alpha$
errors	$1/\sqrt{T}$	β/T	$\frac{1}{\alpha T}$	e^{-T/κ_f}
stepsize	$1/\sqrt{T}$	$\frac{1}{\beta}$	$\frac{1}{\alpha T}$	$\frac{1}{\beta}$

The PL-inequality shows that if the gradient is small for a strongly convex function, then the suboptimality gap is also small. $\|\nabla f(x)\|_2^2 \geq 2\alpha (f(x) - f(x^*))$

In fact, if f is strongly convex then x^* is unique, and the gradient being small also means we are close to the optimizer:

$$\|\nabla f(x)\|_2^2 \geq \alpha^2 \|x - x^*\|_2^2$$

Pf

By strong convexity, $f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\alpha}{2}\|x - x^*\|_2^2$
$= f(x^*) + \frac{\alpha}{2}\|x - x^*\|_2^2$,

so by PL-ineq,

$\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geq f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|_2^2$.

Rearranging, $\|\nabla f(x)\|_2^2 \geq \alpha^2 \|x - x^*\|_2^2$