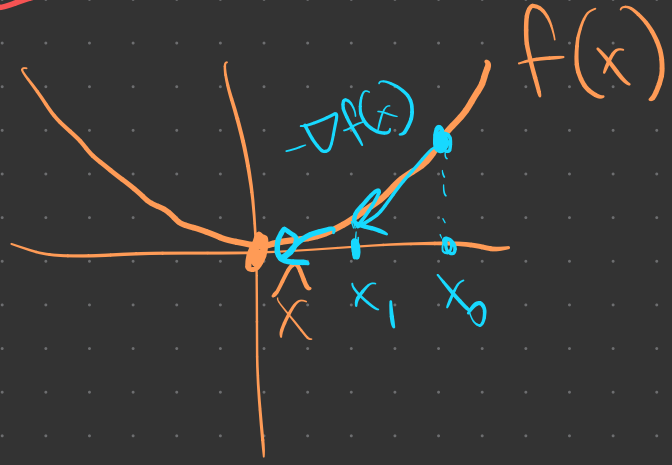
Lecture 1: Basics of Optimization



$$\hat{x} = \arg\min_{x} f(x)$$

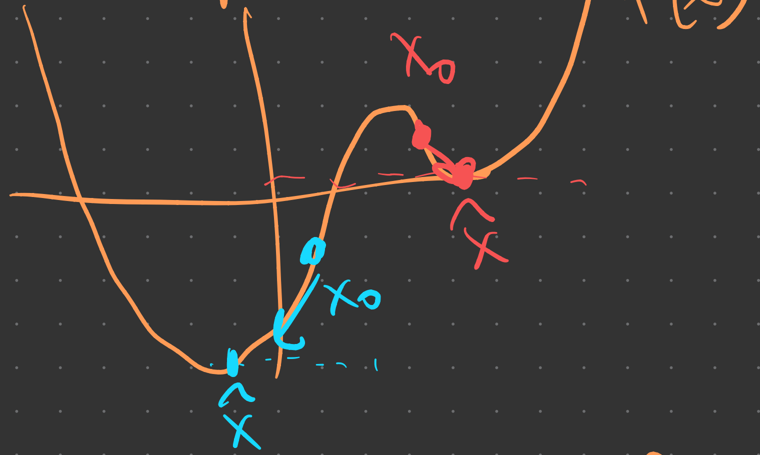
Usually, $x \in \mathbb{R}^m$, n = a lot

Given initial guess x_0, can write

$$x_i = x_{i-1} - \alpha \nabla f(x)_{i-1}$$

$$\longrightarrow \nabla f = 0$$

More complicated:



Recall $x \in \mathbb{R}^m$, maybe $f(x) \in \mathbb{R}^n$, gets challenging fast.

Linear functions:

$f: \mathbb{R}^m \to \mathbb{R}^n$; $f(x) = Ax$ for $A \in \mathbb{R}^{m \times n}$

If mn large, solving $Ax = y$ is very difficult directly $\Longrightarrow$ Constrained Optimization

Take $f(x) = \|Ax-y\|_2^2 := \left(\sqrt{\sum_{i=1}^{n} (Ax-y)_i^2} \right)^2$

$\nabla f(x) = A^T(Ax-y) = 0 \Rightarrow A^TAx - A^Ty = 0 \Rightarrow x = (A^TA)^{-1}A^Ty$

also may be hard to invert...

$x_k = x_{k-1} - \alpha \underbrace{A^T(Ax_k - y)}_{\nabla f(x_k)}$

Note: $A \in M_n(\mathbb{C}) \Rightarrow f(x) = (Ax-y)^H(Ax-y)$, where $A^H = (A^T)^*$ is the conjugate transpose of A. One can derive the relations

$$x_k = x_{k-1} - \alpha A^H(Ax_k - y)$$

$$x = (A^HA)^{-1}A^Hy$$

We may want more sophisticated optimization using probability

Let X be a random variable, with known observation y.

Assume Gaussian prior: $P(y|x) \propto e^{-\frac{(Ax-y)^2}{\sigma^2}}$.

Maximum a posteriori (MAP) estimator: $\boxed{\max P(x|y)}$

<u>Bayes Thm</u>: $p(x|y) = \dfrac{p(y|x)\cancel{p(y)}^{1}}{p(x)}$

$$\Rightarrow \log p(x|y) = \log p(y|x) - \log p(x)$$

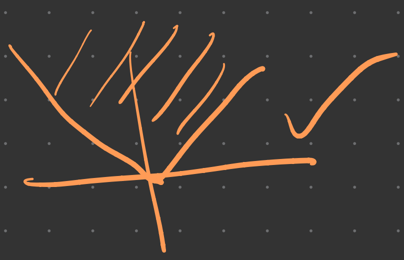
$$= -\frac{1}{\sigma^2}(Ax-y)^2 - \frac{1}{\sigma^2}x^2$$

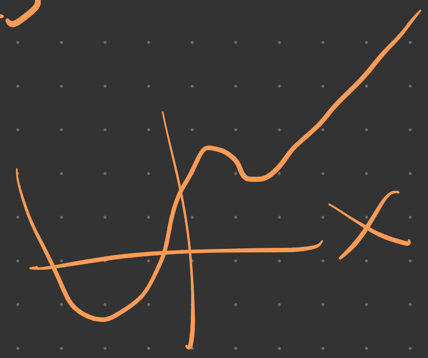
$\longrightarrow \min[-\log p(x|y)] = \frac{1}{\sigma^2}(Ax-y)^2 + \frac{1}{\sigma^2}x^2$ looks like $\|Ax-y\|_2^2 + R(x)$

Characteristics of $f(x)$ that help optimization:

$f(x)$ differentiable: Can analytically compute $\nabla f(x)$

$f(x)$ convex: local min = global min





$$f(tx_0) + f((1-t)x_1) \leq t f(x_0) + (1-t) f(x_1)$$

The game is to automatically find $-\nabla f(x)$, which is why autograd/backprop exists

Exercises:

1. $f(x) = \|Ax - y\|_2^2 + \lambda\|x\|_2^2$. What is the closed-form expression for $\hat{x} = \underset{x}{\arg\min}\, f(x)$?

2. Let $A: \mathbb{R}^m \to \mathbb{R}^n$. What is the shape of A^TA? Is A^TA easier to invert if $m \ll n$ or $m \gg n$?

3. Consider $f(x): [0,1] \to \mathbb{R}$.

   a) Give an example of f that is diff'ble but not convex.

   b) " " convex but not diff'ble

For each case, show a visual example of the gradient descent optimization process.

Lecture 2: Function Spaces

V is a vector space (over $\mathbb{R}$) if $\forall v, w \in V$, $\alpha \in \mathbb{R}$, we have:

① $v + w \in V$, ② $\alpha v \in V$ s.t. $\alpha(v+w) = \alpha v + \alpha w \in V$.

Such elements v, w are called vectors. Usually think of $v \in \mathbb{R}^n$, but this lecture will go over more exotic spaces.

$B = \{v_1, \ldots, v_n\} \subseteq V$ is a basis if $\forall w \in V$, $\exists! \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ st

$w = \alpha_1 v_1 + \cdots + \alpha_n v_n$. $n = \dim V$ is the dimension of V.

Fact: If B_1, B_2 are bases of V, then $\#B_1 = \#B_2 = n$.

Sps $B_1 = \{v_1, \ldots, v_n\}$, $B_2 = \{w_1, \ldots, w_n\}$. Then by defn $\exists!$

$\hookrightarrow \alpha_{11}, \ldots, \alpha_{1n}$ st $v_1 = \alpha_{11} w_1 + \alpha_{12} w_2 + \cdots + \alpha_{1n} w_n$

$\hookrightarrow \alpha_{21}, \ldots, \alpha_{2n}$ st $v_2 = \alpha_{21} w_1 + \alpha_{22} w_2 + \cdots + \alpha_{2n} w_n$

$\vdots$

$v_n = \alpha_{n1} w_1 + \alpha_{n2} w_2 + \cdots + \alpha_{nn} w_n$

1 unique way to convert $B_1 \to B_2$

Can write this more succinctly as $\vec{v} = \Phi \vec{w}$ where we know Φ is invertible.

Fact: every vector space has a basis.!!

Linear Transform $T: V \to W$ is a function where

① $T(v+w) = T(v) + T(w)$ ② $T(\alpha v) = \alpha T(v)$

Note:

Can always write $T = \Phi_W T' \Phi_V$, applying transformation in arbitrary basis.

Let's do some examples:

<u>Ex 1</u> $V = \mathrm{Span}_{\mathbb{R}}\{\sin(ax) : a \in \mathbb{Z}\}$, $W = \mathrm{Span}_{\mathbb{R}}\{\cos(ax) : a \in \mathbb{Z}\}$

$\frac{d}{dx} : V \to W$ is a linear transform. $\dim V = \dim W = \infty$.

<u>Ex 2</u> $V = \mathrm{Span}_{\mathbb{R}}\{1, x, \dots, x^2\}$.

① What is Φ_V, changing to the basis $\{1, x+1, (x+1)^2\}$?

② What is the matrix from of $\frac{d}{dx}$ in the orijinal basis?

A normed vector space $(V, |\cdot|)$ has a norm $|\cdot|$ that satisfies:

i) $|x| = 0 \iff x = 0 \in V$

ii) $|x+y| \le |x| + |y|$

iii) $|\lambda x| = |\lambda||x|$

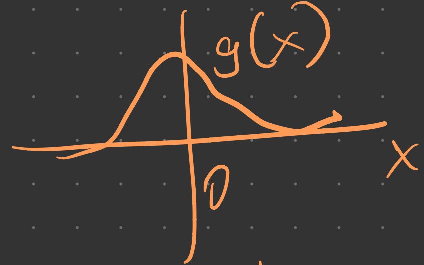
Fact: $|x \cdot y| \le |x||y|$ (Cauchy-Schwartz)

Def: Let $f: \Omega \to \mathbb{R}$, $f \in V$. $\|f\|_p := \left(\int_\Omega f^p dx\right)^{1/p}$ be denoted as the L^p norm of f.

Remark: In ML, you will often see L^1, L^2, sometimes L^0.

L^2 behaves nicely, L^1 promotes sparsity, L^0 is even better for sparsity.

<u>Convolution</u>: $(f \star g)(x) = \int_{\mathbb{R}} f(a)\, g(x-a)\, da$



(usually)

<u>Fourier Transform</u>:

$$\hat{f}(k) = \int_{\mathbb{R}} f(x)\, e^{-2\pi i x k}\, dx \quad , \quad \check{f}(x) = \int_{\mathbb{R}} f(k)\, e^{+2\pi i x k}\, dk$$

<u>Facts</u>

① $\|f(x)\|_2 = \|\hat{f}(k)\|_2$ (Plancherel/Parseval)

② $\widehat{(f \star g)} = \hat{f}\hat{g}$ (Convolution-Multiplication)

③ $f(x) = Ae^{-x^2} \Longleftrightarrow \hat{f}(k) = Be^{-k^2}$

Exercises

1) Show that the change of basis $\Phi : V \to V$ is a linear transform.

2) V as in Ex 2. what is the vector space W st $\int dx\,(V) = W$?

3) Define $\delta(x) = \lim\limits_{\sigma \to \infty} \frac{1}{1+\sigma^2 x^2}$, $\int_{\mathbb{R}} f(x)\delta(x) = f(0)$.

What is $\hat{\delta}(x)$? I.e. what is Fourier transform of a δ?

Lecture 3: Linear Algebra

$$M = [M]_{ij} = \begin{bmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & & \\ M_{n1} & \cdots & M_{nn} \end{bmatrix}$$

$$[M+N]_{ij} = [M]_{ij} + [N]_{ij}$$

$$[MN]_{ij} = \begin{bmatrix} \quad & & \\ & & \end{bmatrix} = \underbrace{\begin{bmatrix} \quad & & \\ & & \end{bmatrix}}_{M}\underbrace{\begin{bmatrix} \quad & & \\ & & \end{bmatrix}}_{N}$$

$$[MN]_{ij} = \sum_{k} M_{ik} N_{kj}$$

Tensor Product: $v \otimes w = \begin{bmatrix} v_1 w_1 & \cdots & v_n w_1 \\ \vdots & & \vdots \\ v_n w_1 & & v_n w_n \end{bmatrix}$

"Outer Product"

Iterative linear solving: RREF

$$\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} x = \begin{bmatrix}1\\ 2\end{bmatrix}$$

$$\downarrow$$

$$\left[\begin{array}{cc|c}1 & 2 & 1\\ 3 & 4 & 2\end{array}\right]$$

$① - 3② \leftarrow ②$

$$\downarrow$$

$$\left[\begin{array}{cc|c}1 & 2 & 1\\ 0 & -2 & -1\end{array}\right]$$

$② \to -\frac{②}{2}$

$$\downarrow$$

$$\left[\begin{array}{cc|c}1 & 2 & 1\\ 0 & 1 & \frac{1}{2}\end{array}\right]$$

$① \to ① - 2②$

$$\downarrow$$

$$\left[\begin{array}{cc|c}1 & 0 & 0\\ 0 & 1 & \frac{1}{2}\end{array}\right]$$

$$\Longrightarrow \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} x = \begin{bmatrix}0\\ \frac{1}{2}\end{bmatrix}$$

$$\Longrightarrow x = \begin{bmatrix}0\\ 1/2\end{bmatrix}$$

exactly one solution

$A = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}$ is invertible

☆ Show $v \otimes w$ is not invertible for $v, w \in \mathbb{R}^2$

Determinant

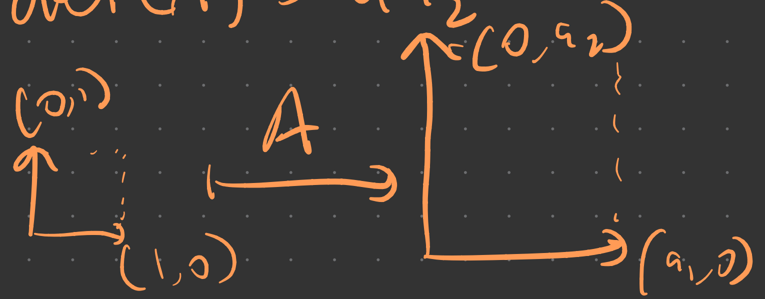
$$\det \begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix} = M_{11}M_{22} - M_{12}M_{21}$$

Invertible $\Longleftrightarrow \det M \neq 0$

$$\det(AB) = \det(A)\det(B)$$

If $A = \begin{bmatrix}a_1 & 0\\ 0 & a_2\end{bmatrix}$,

$$\det(A) = a_1 a_2$$



Eigenvectors/values

$$Mx = \lambda x$$

matrix vector # vector

$$\longrightarrow Mx = \lambda I x \qquad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$(M - \lambda I)x = 0 \leftarrow \text{matrix}$$

$$\det(M - \lambda I) = 0 \leftarrow \mathbb{R}$$

↳ characteristic polynomial of M

<u>Fact</u>: If $A \in M_n(\mathbb{C})$, char. poly. of A has n complex roots.

Diagonalization:

A matrix $T \in M_n(\mathbb{R})$ is diagable iff

$T = \Lambda S \Lambda^{-1}$, where Λ is a change of basis, and S is a diagonal matrix, i.e.

$$S = \begin{bmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_n \end{bmatrix}.$$

<u>Spectral Theorem</u>: Suppose $A \in M_n(\mathbb{R})$ st $A = A^T$. Then:

① A is diag'ble

② A has real eigenvalues.

Now, what if $A \in M_{m\times n}(\cancel{\mathbb{R}})$? Singular Value Thm!
$\mathbb{C}$

<u>SVD</u>: <u>Every</u> matrix $A \in M_{m\times n}(\mathbb{C})$ can be written as

$$A = U\Sigma V^H, \text{ st } UU^H = VV^H = I,$$

$$\Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix} \text{ w/ } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

Important! $\operatorname{rk}(A) = \#\{\sigma_n \geq 0\}$ where $A = U\begin{bmatrix}\sigma_1 & & 0\\ & \ddots & \\ 0 & & \sigma_n\end{bmatrix}V^H$.

This decomposition is powerful, since it "orders the components of A from most important to least important." (See Exercise 3)

Application: Matrix Approximation

$$\underset{\operatorname{rk} B \leq k}{\arg\min} \|A-B\|_F^2 = U\begin{bmatrix}\sigma_1 & & & 0\\ & \ddots & & \\ 0 & & \sigma_k & \\ & & & 0\end{bmatrix}V^H$$

(Eckart-Young)

i.e. SVD enables optimal low-rank approximation.

Exercises

① Show that $\det(A^{-1}) = \dfrac{1}{\det A}$, for invertible A.

② Show that $u \otimes v \in M_n(\mathbb{R})$ is not invertible, for $u, v \in \mathbb{R}^n$.

③ Show that $A = U \Sigma V^H$, $\Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}$ can be written as $A = \sum_{i=1}^{n} \sigma_i \, u_i \otimes v_i$, for $u_i, v_i \in \mathbb{C}^n$.