Nonlinear Optimization in Machine Learning

Lecture 1 Introduction & Foundations

Why nonlinear optimization?

motivated by Machine Learning Applications

$$\mathcal{D} = \{(a_j, y_j), j = 1, 2, \ldots, m\}$$

"learn" $\phi = \phi(a; x)$

"loss" $$\mathcal{L}_D(x) = \sum_{j=1}^{m} \ell(a_j, y_j; x) = \sum_{j=1}^{m} \ell(\phi(a_j; x), y_j)$$

$$x^* = \min_{x \in U} \mathcal{L}_D(x)$$

Example 1 Least Squares

$$\min_x \frac{1}{2m} \sum_{j=1}^{m} (a_j^T x - y_j)^2 = \frac{1}{2m} \|Ax - y\|_2^2$$

"regularization"

$$\min_{x} \frac{1}{2m} \|Ax - y\|_2^2 + \lambda \|x\|_2^2 \qquad (\lambda > 0)$$

(Tikhonov regularization)

$$\min_{x} \frac{1}{2m} \|Ax - y\|_2^2 + \lambda \|x\|_1$$

(LASSO: Least Absolute Shrinkage and Selection Operator)

<u>Example 2</u> Matrix Completion

A_j is $n \times p$ and X is $n \times p$

$$\min_{X} \frac{1}{2m} \sum_{j=1}^{m} (\langle A_j, X \rangle - y_j)^2$$

where $\langle A, B \rangle = tr(A^T B)$

$$\min_{X} \frac{1}{2m} \sum_{j=1}^{m} (\langle A_j, X \rangle - y_j)^2 + \lambda \|X\|_*$$

$\|X\|_*$ = sum of |singular values| of $X = tr\sqrt{X^T X}$

= nuclear norm

$L \in \mathbb{R}^{n \times r}$ and $R \in \mathbb{R}^{p \times r}$ $\quad r \ll \min(n, p)$

$$\min_{L,R} \frac{1}{2m} \sum_{j=1}^{m} \left(\langle A_j, LR^T \rangle - y_j \right)^2$$

Example 3 Nonnegative matrix factorization

$$\min_{L,R} \|LR^T - Y\|_F^2 \quad , \; L \geqslant 0, \; R \geqslant 0$$

$Y \in \mathbb{R}^{n \times p}$ $\quad L \in \mathbb{R}^{n \times r}$ $\quad R \in \mathbb{R}^{p \times r}$

Example 4 Sparse inverse covariance estimation

Sample covariance matrix $\quad S = \frac{1}{m-1} \sum_{j=1}^{m} a_j a_j^T$

$S^{-1} = X$

$$\min_{\substack{X \in \text{Symmetric } \mathbb{R}^{n \times n} \\ X \succeq 0}} \langle S, X \rangle - \log\det |X| + \lambda \|X\|_1$$

"Graphical LASSO"

$\|X\|_1 = \sum_{i,l=1}^{n} |X_{il}|$

Example 5 Sparse PCA

PCA = Principle component analysis

$$\max_{v \in \mathbb{R}^n} v^T S v \quad \text{s.t } \|v\|_2 = 1, \ \|v\|_0 \le k$$

"Sparse" via R $\quad M = vv^T$

$$\max_{M \in \text{Symmetric } \mathbb{R}^{n \times n}} \langle S, M \rangle \quad \text{s.t. } M \succeq 0, \ \langle I, M \rangle = 1, \ \|M\|_1 \le R$$

Example 6 SVM (Support Vector Machine)

$a_j \in \mathbb{R}^n \quad y_j \in \{-1, 1\}$

seek $x \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ s.t. $a_j^T x - \beta \ge 1$ if $y_j = 1$
$a_j^T x - \beta \le -1$ if $y_j = -1$

$$H(x, \beta) = \frac{1}{m} \sum_{j=1}^{m} \max\left(1 - y_j(a_j^T x - \beta), 0\right)$$

Example 7 Neural Network

"activation function" σ

$$a_j^{\ell} = \sigma(W^{\ell} a_j^{\ell-1} + g^{\ell}), \quad \ell = 1, 2, \ldots, D$$

"weight" $w = (W^1, g^1, W^2, g^2, \ldots, W^D, g^D)$

$$L(w, X) = \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{\ell=1}^{M} y_{j\ell} (x_{[\ell]}^T a_j^D(w)) - \log \left(\sum_{\ell=1}^{M} \exp(x_{[\ell]}^T a_j^D(w)) \right) \right]$$

"logistic regression"

Fundations of Optimization

$$f : D \subset \mathbb{R}^n \to \mathbb{R}$$

"local minimizer" "global minimizer"

"strict local minimizer"

"isolated local minimizer"

Constrained Optimization Problem

$$\min_{x \in \Omega} f(x)$$

where $\Omega \subset D \subset \mathbb{R}^n$ is a closed set

"local solution" "global solution"

relation with unconstrained $\min f(x) + I_\Omega(x)$

$$I_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{otherwise} \end{cases}$$

Convexity

"convex set" Ω $\quad x, y \in \Omega \Rightarrow (1-\alpha)x + \alpha y \in \Omega \quad \forall \alpha \in [0,1]$

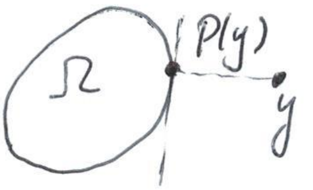
"Supporting hyperplane for Ω at $\bar{x} \in \Omega$"

is defined by $g \in \mathbb{R}^n$ $g \neq 0$ s.t

$$g^T(x - \bar{x}) \leq 0 \quad \text{for all } x \in \Omega$$

Projection Operator $P: \mathbb{R}^n \to \Omega$

$$P(y) = \arg\min_{z \in \Omega} \|z - y\|_2^2$$



Convex function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$

$$\phi((1-\alpha)x + \alpha y) \leq (1-\alpha)\phi(x) + \alpha\phi(y)$$

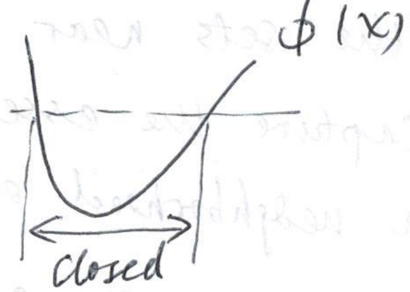
$$\forall x, y \in \mathbb{R}^n, \quad \forall \alpha \in [0,1]$$

"effective domain" $= \{x \in \Omega . \phi(x) < +\infty\}$

"epigraph" $=$ epi $\phi := \{(x,t) \in \Omega \times \mathbb{R} : t \geq \phi(x)\}$

"proper convex function"

"closed proper convex function"

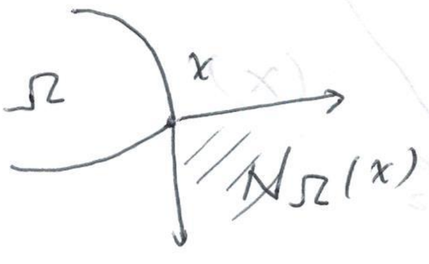


Definition "Normal Cone"

$\Omega \subset \mathbb{R}^n$ is convex set

$N_\Omega(x)$ = normal cone at $\forall x \in \Omega$

$$= \{d \in \mathbb{R}^n : d^T(y-x) \leq 0 \text{ for all } y \in \Omega\}$$



<u>Theorem</u> If Ω_i, $i=1,2,\dots,m$ are convex sets and $\Omega = \bigcap_{i=1,2,\dots m} \Omega_i$, then $\forall x \in \Omega$

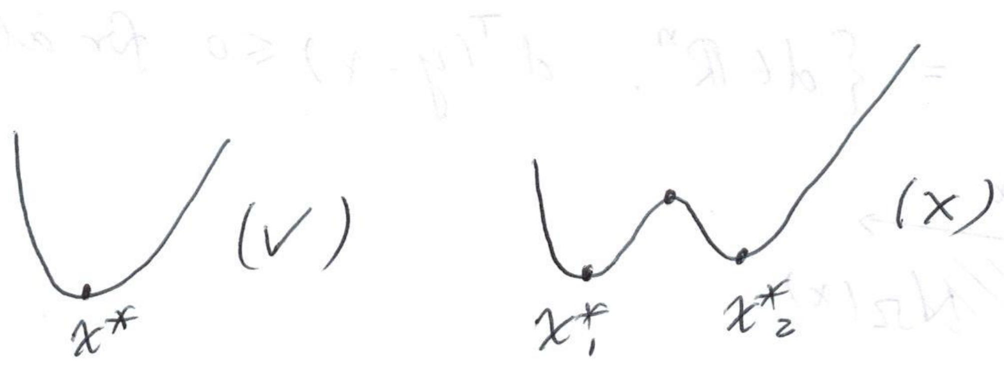
$$N_\Omega(x) \supset N_{\Omega_1}(x) + N_{\Omega_2}(x) + \dots + N_{\Omega_m}(x)$$

For "=" in the above we need <u>constraint qualifications</u>: a linear approximation of the sets near the point in question needs to capture the essential geometry of the set itself in a neighborhood of the point.

<u>Theorem</u> If f is convex and Ω closed convex then for $\min_{x\in\Omega} f(x)$ we have

(a) any local solution is a global solution

(b) the set of global solutions form a convex set.



Important quantities

"modulus of continuity m for strongly convex ϕ"

$m > 0$ $\quad \forall x, y \in$ domain of ϕ

$$\phi((1-\alpha)x + \alpha y) \le (1-\alpha)\phi(x) + \alpha\phi(y) - \frac{1}{2} m \alpha (1-\alpha) \|x-y\|_2^2 \quad (*)$$

<u>Theorem</u> (Taylor's formula)

$f: \mathbb{R}^n \to \mathbb{R}$ continuously differentiable

$x, p \in \mathbb{R}^n$

$$f(x+p) = f(x) + \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma$$

$$f(x+p) = f(x) + \nabla f(x+\gamma p)^T p \quad \text{for some } \gamma \in (0,1)$$

$f: \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p) p \, d\gamma$$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+\gamma p) p$$

for some $\gamma \in (0,1)$

Lipschitz constant L for ∇f

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \qquad (**)$$

for all $x, y \in dom(f)$

Theorem

(1) If f is continuously differentiable and conve

then $$f(y) \geq f(x) + (\nabla f(x))^T (y-x)$$

for $\forall x, y \in dom(f)$

(2) If f is differentiable and $\boxed{\text{strongly}}$ convex then

$$f(y) \geq f(x) + (\nabla f(x))^T (y-x) + \frac{m}{2}\|y-x\|^2$$

for $\forall x, y \in dom(f)$

(3) If ∇f is uniformly Lipschitz continuous with Lipschitz constant L and f is convex then

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

for $\forall x, y \in dom(f)$

<u>Proof</u>. (1) $\partial f(x) = \{\nabla f(x)\}$

$z \to x$ $\quad f(z) \geq f(x) + (\nabla f(x))^T (z-x)$

$$\begin{array}{c} f(z) \\ \wedge\!\!| \\ \alpha f(y) + (1-\alpha) f(x) \end{array} \quad (0<\alpha<1)$$

So $\quad \alpha f(y) \geq \alpha f(x) + (\nabla f(x))^T (z-x)$

$\Rightarrow \quad f(y) \geq f(x) + (\nabla f(x))^T \underbrace{\left(\frac{z-x}{\alpha}\right)}_{= (y-x)}$

(2) follows (*)

(3) By Taylor expansion

$$f(y) - f(x) - (\nabla f(x))^T (y-x) = \int_0^1 [\nabla f(x+\gamma(y-x)) - \nabla f(x)]^T (y-x)\, d\gamma$$

$$\leq \int_0^1 \|\nabla f(x+\gamma(y-x)) - \nabla f(x)\| \cdot \|y-x\|\, d\gamma$$

$$\leq \int_0^1 L\gamma \|y-x\|^2 d\gamma = \frac{L}{2} \|y-x\|^2$$

<u>Theorem</u> $f \in C^2(\mathbb{R}^n)$

f is strongly convex with modulus of corexity m

$\Leftrightarrow \quad \nabla^2 f(x) \succeq mI$ for all x

∇f is Lipschitz continuous with Lipschitz constant

$L \Leftrightarrow \quad \nabla^2 f(x) \preceq LI$ for all x

<u>Theorem</u> If f is dofferentiable and strongly convex with modulus of continuity m Then minimizer x^* of f exists and is unique.

<u>Key to the proof</u> ①. $\{x \mid f(x) \le f(x^0)\}$ for any x^0 is colosed and bounded

② x^* is unique.

<u>Theorem</u> f is convex, ∇f with Lipschitz constant L then $\forall\ x, y \in dom(f)$

$$f(x) + (\nabla f(x))^T (y-x) + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$$

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))^T (x-y) \leq L\|x-y$$

If in addition f is strongly convex and with modulus of convexity m, unique minimizer x^*

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

$\forall\ x, y \in dom(f)$

<u>Proof</u>. Define $\phi(y) = f(y) - (\nabla f(x))^T y$

ϕ is convex $\quad \nabla \phi(y) = \nabla f(y) - \nabla f(x)$

$$\nabla \phi(x) = \nabla f(x) - \nabla f(x) = 0$$

so x is a minimizer of ϕ

so $$\phi(x) \leq \phi\left(y - \frac{1}{L} \nabla \phi(y)\right)$$

$$\leq \phi(y) + (\nabla \phi(y))^T \left[-\frac{1}{L} \nabla \phi(y)\right] + \frac{L}{2} \|(-\frac{1}{L}) \nabla \phi(y)\|$$

$$= \phi(y) - \frac{1}{2L} \|\nabla \phi(y)\|^2$$

so $\quad f(x) - (\nabla f(x))^T x$

$$\leq f(y) - (\nabla f(x))^T y - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

i.e. $f(y) \geq f(x) + (\nabla f(x))^T (y-x) + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|$

Same way $f(x) \geq f(y) + (\nabla f(y))^T (x-y) + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|$

$$\Rightarrow \left[(\nabla f(x))^T - (\nabla f(y))^T \right] (x-y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|$$

Finally by (*)

$$f(y) - f(x) \geq (\nabla f(x))^T (y-x) + \frac{m}{2} \|y-x\|^2$$

$$= \frac{1}{2m} \|\nabla f(x)\|^2 + (\nabla f(x))^T (y-x) + \frac{m}{2} \|y-x\|^2 - \frac{1}{2m} \|\nabla f(x)\|^2$$

$$= \frac{m}{2} \left\| y - x + \frac{1}{m} \nabla f(x) \right\|^2 - \frac{1}{2m} \|\nabla f(x)\|^2$$

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

#

$$\|\nabla f(x)\|^2 \geqslant 2m\,[f(x) - f^*], \quad m > 0$$

"generalized strong convexity condition"

"quadratic surrogate"

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|$$

Optimality conditions for smooth unconstrained problems

Theorem (Necessary Conditions for smooth unconstrained optimization)

(a). f is continuously differentiable, x^* — local minimizer of $\min\limits_{x \in \mathbb{R}^n} f(x)$ then $\nabla f(x^*) = 0$ (first-order necessary condition)

(b). f is twice continuously differentiable, x^* — local minimizer of $\min\limits_{x \in \mathbb{R}^n} f(x)$ then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite

(second-order necessary condition)

When f is convex, first-order necessary condition became sufficient

<u>Theorem</u> f is continuously differentiable and convex

$\nabla f(x^*) = 0 \Rightarrow x^*$ is global minimizer of $\min_{x \in \mathbb{R}^n} f(x)$

f is strongly convex $\Rightarrow x^*$ is unique

<u>Key</u> $f(y) \geq f(x^*) + (\nabla f(x^*))^T (y - x^*)$
$= f(x^*)$

When f is non-convex

<u>Theorem</u> (Second-order sufficient condition)

If f is twice continuously differentiable and that for some x^* we have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite

Then x^* is a strict local minimizer of

$\min_{x \in \mathbb{R}^n} f(x)$.

Optimizality conditions for smooth constrained problem = nonsmooth problems

$$\min_{x \in \mathbb{R}^n} \left[f(x) + I_\Omega(x) \right]$$

<u>Theorem</u> Let Ω be closed and convex in $\mathbb{R}^n$

Let f be convex and differentiable

x^* is a minimizer of $\min_{x \in \mathbb{R}^n} [f(x) + I_\Omega(x)]$

$\Leftrightarrow \quad -\nabla f(x^*) \in N_\Omega(x^*)$

$$\boxed{\partial I_\Omega(x^*) = N_\Omega(x^*)} \quad (\text{key})$$

$\forall d \in \partial I_\Omega(x^*) \qquad x^* \in \Omega$

$I_\Omega(x) \geq I_\Omega(x^*) + d^T(x - x^*)$

so $\quad d^T(x - x^*) \leq 0$ of $x, x^* \in \Omega$