COMP7180: Quantitative Methods for Data Analytics and Artificial Intelligence

Lecture 6: Convex Optimization: Algorithms

Jun Qi

Research Assistant Professor in Computer Science @ Hong Kong Baptist University

Affiliated Associate Professor in Electronic Engineering @ Fudan University

香港浸會大學 HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系

29/10/2024

CVX Course Scope

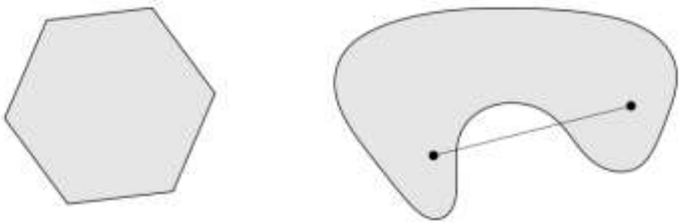
- Convex sets and convex functions
- Convex function properties: 1st and 2nd characterizations
- Lagrangian Multiplier Method
- Conjugate Functions and Dual Norm
- First-order Methods (GD and SGD)

Convex sets

Convex set: $C \subseteq \mathbb{R}^n$ such that

$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \le t \le 1$$

In words, line segment joining any two elements lies entirely in set



Convex combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: any linear combination

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

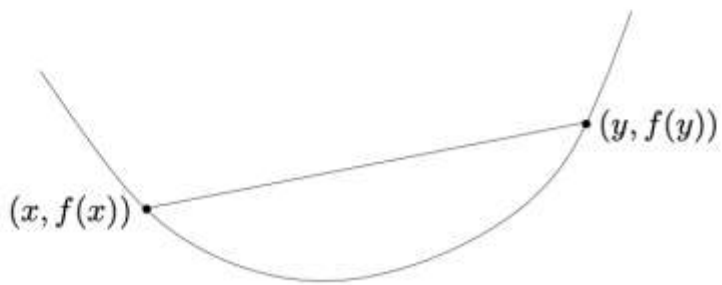
with $\theta_i \geq 0$, $i = 1, \ldots, k$, and $\sum_{i=1}^k \theta_i = 1$. Convex hull of a set C, $\mathrm{conv}(C)$, is all convex combinations of elements. Always convex

Convex functions

Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for } 0 \leq t \leq 1$$

and all $x, y \in \mathrm{dom}(f)$



In words, function lies below the line segment joining $f(x), f(y)$

Concave function: opposite inequality above, so that

$$f \text{ concave} \iff -f \text{ convex}$$

First-order Convexity Condition

Theorem 1. Assume that f($\mathbf{x}$) is differentiable, then f($\mathbf{x}$) is convex

if and only if

the domain C is convex and $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})$.

The proof can be found in Proposition 4 in
https://wiki.math.ntnu.no/_media/tma4180/2016v/note2.pdf

We use the linear function to check the result.

$f(\mathbf{x}) = \mathbf{Ax}+\mathbf{b}$, then $\nabla f(\mathbf{x})^T = \mathbf{A}$.

$f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})=\mathbf{Ax}+\mathbf{b}+\mathbf{A}(\mathbf{y}-\mathbf{x})=\mathbf{Ay}+\mathbf{b}$

So $f(\mathbf{y})=f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})$

Second-order Convexity Condition

Theorem 2. Assume that f(**x**) is twice differentiable, then f(**x**) is convex

if and only if

the domain C is convex and the Hessian Matrix **H**(**x**) is positive semi-definite.

The proof can be found in Proposition 7 in https://wiki.math.ntnu.no/_media/tma4180/2016v/note2.pdf

- What is positive semi-definite matrix **M**?

Positive semi-definite matrix **M** is a nxn sysmetric matrix **M** = $\mathbf{M}^{\mathrm{T}}$ and for any real n-dimensional vector **z,** $\mathbf{z}^{\mathrm{T}}\mathbf{M}\mathbf{z} \geq 0$.

Positive Semi-definite Matrix (1/3)

- Definition: A symmetric matrix A (i.e., $A = A^T$) is said to be **positive semidefinite**, if for any non-zero vector x, the following condition holds:

$$x^T A x \geq 0.$$

- Key properties:

1. **Eigenvalues**: All eigenvalues of the positive semi-definite matrix are non-negative.

   If λ is an eigenvalue of A, then $\lambda \geq 0$.

2. **Leading Principal Minors:**

   For a 3x3 matrix: $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

   Leading Principal Minors are:

$$\det(A_1) = |a_{11}| \qquad \det(A_2) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \qquad \det(A_3) = \det(A)$$

Positive Semi-definite Matrix (2/3)

- Example 1:

Consider matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

Leading Principal Minors:

First order: $\det(A_1) = |1| = 1 > 0$

Second order: $\det(A_2) = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} = 0$

Since all leading principal minors are ≥ 0, A is positive semidefinite.

- Example 2:

Consider matrix $A = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$

Leading Principal Minors:

First order: $\det(A_1) = |4| = 4 > 0$

Second order: $\det(A_2) = \begin{vmatrix} 4 & 2 \\ 2 & 1 \end{vmatrix} = 4(1) - 2(2) = 0$

Since all leading principal minors are ≥ 0, A is positive semidefinite.

Positive Semi-definite Matrix (3/3)

- Example 3

Let's identify if matrix

$$A = \begin{bmatrix} 4 & 2 & 5 \\ 2 & 1 & 3 \\ 5 & 3 & 6 \end{bmatrix}$$

is positive semidefinite using leading principal minors.

For a matrix to be positive semidefinite, all leading principal minors must be ≥ 0.

Let's calculate each leading principal minor:

First leading principal minor (1×1):

$$\det(A_1) = |4| = 4 > 0$$

Second leading principal minor (2×2):

$$\det(A_2) = \begin{vmatrix} 4 & 2 \\ 2 & 1 \end{vmatrix} = 4(1) - 2(2) = 4 - 4 = 0 \geq 0$$

Third leading principal minor (3×3):

$$\begin{aligned} \det(A_3) &= \begin{vmatrix} 4 & 2 & 5 \\ 2 & 1 & 3 \\ 5 & 3 & 6 \end{vmatrix} \\ &= 4\begin{vmatrix} 1 & 3 \\ 3 & 6 \end{vmatrix} - 2\begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix} + 5\begin{vmatrix} 2 & 1 \\ 5 & 3 \end{vmatrix} \\ &= 4(6-9) - 2(12-15) + 5(6-5) \\ &= 4(-3) - 2(-3) + 5(1) \\ &= -12 + 6 + 5 \\ &= -1 < 0 \end{aligned}$$

Since the third leading principal minor is negative, this matrix is NOT positive semidefinite.

Optimization terminology

Reminder: a convex optimization problem (or program) is

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \ i = 1, \dots, m \\ & Ax = b \end{aligned}$$

where f and g_i, $i = 1, \dots, m$ are all convex, and the optimization domain is $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i)$ (often we do not write D)

- f is called criterion or objective function
- g_i is called inequality constraint function
- If $x \in D$, $g_i(x) \le 0$, $i = 1, \dots, m$, and $Ax = b$ then x is called a feasible point
- The minimum of $f(x)$ over all feasible points x is called the optimal value, written $f^\star$

Rewriting constraints

The optimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \dots, m \\ & Ax = b \end{array}$$

can be rewritten as

$$\min_{x} \ f(x) \quad \text{subject to} \quad x \in C$$

where $C = \{x : g_i(x) \leq 0, \ i = 1, \dots, m, \ Ax = b\}$, the feasible set. Hence the latter formulation is completely general

With I_C the indicator of C, we can write this in unconstrained form

$$\min_{x} \ f(x) + I_C(x)$$

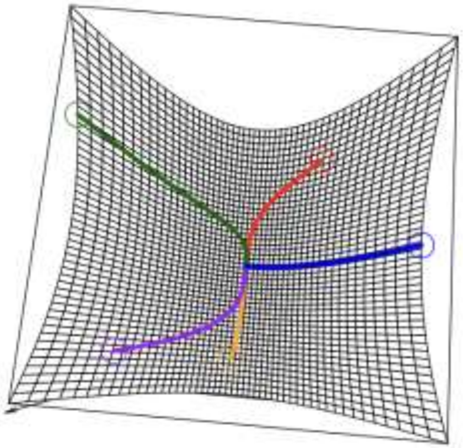
Local minima are global minima

For a convex problem, a feasible point x is called locally optimal is there is some $R > 0$ such that

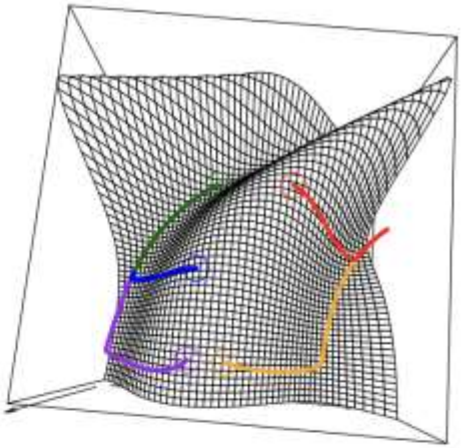
$$f(x) \leq f(y) \quad \text{for all feasible } y \text{ such that } \|x - y\|_2 \leq R$$

Reminder: for convex optimization problems, local optima are global optima

Proof simply follows from definitions



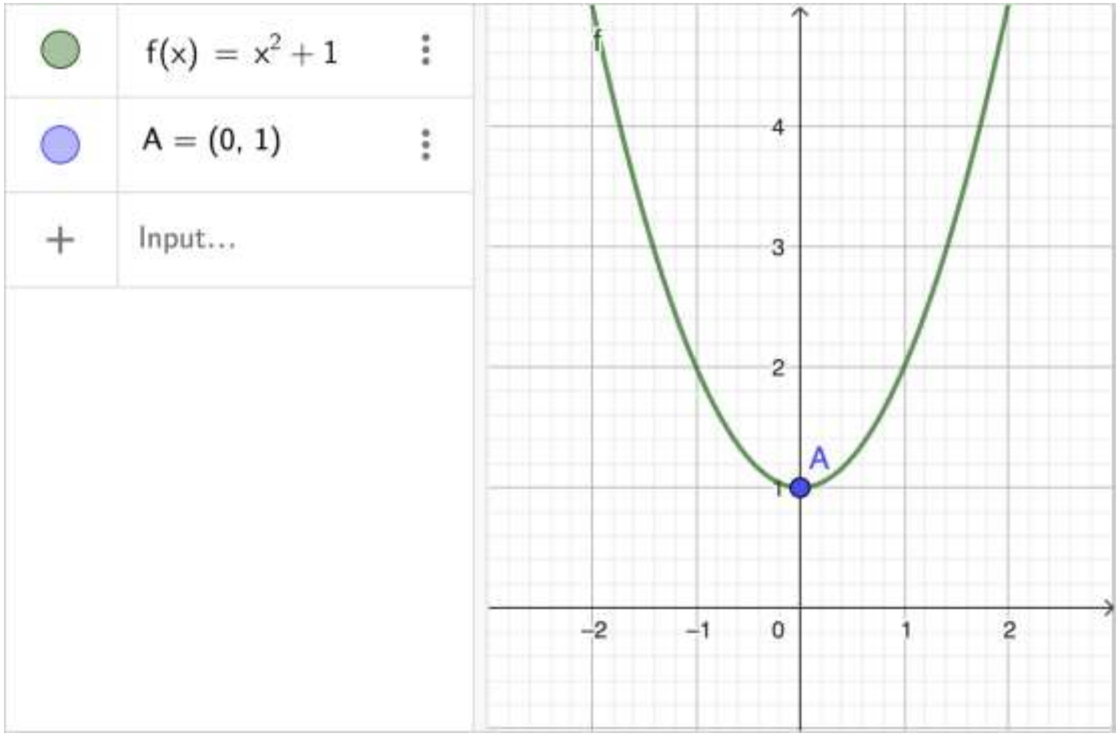
Convex



Nonconvex

Local Minima are Global Minima

- Example: $f(x) = x^2 + 1$



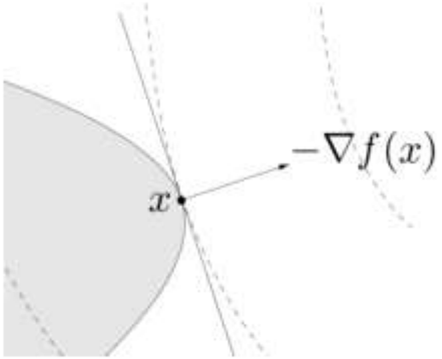
First-order Optimality Condition

For a convex problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \qquad s.t., \boldsymbol{x} \in C$$

and differential *f*, a feasible point $\boldsymbol{x}$ is optimal if and only if

$$\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In other words, all feasible directions from $\boldsymbol{x}$ are aligned with the gradient $\nabla f(\boldsymbol{x})$

Important special case: if $C = R^n$(unconstrained optimization), then optimality condition reduces to $\nabla f(\boldsymbol{x}) = 0$

Example: quadratic minimization

Consider minimizing the quadratic function

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c$$

where $Q \succeq 0$. The first-order condition says that solution satisfies

$$\nabla f(x) = Qx + b = 0$$

- if $Q \succ 0$, then there is a unique solution $x = -Q^{-1}b$
- if Q is singular and $b \notin \mathrm{col}(Q)$, then there is no solution (i.e., $\min_x \ f(x) = -\infty$)
- if Q is singular and $b \in \mathrm{col}(Q)$, then there are infinitely many solutions

$$x = -Q^+ b + z, \quad z \in \mathrm{null}(Q)$$

where Q^+ is the pseudoinverse of Q

Equality-constrained Minimization

Consider the equality-constrained convex optimization:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \qquad s.t., \mathrm{A}\boldsymbol{x} = \mathbf{b}$$

with f differentiable. Let's Lagrange multiplier optimality condition:

$$\nabla f(\boldsymbol{x}) + A^T\boldsymbol{\lambda} = 0, \text{ for a vector } \boldsymbol{\lambda}$$

Proof: the constrained convex optimization problem can be reformulated as:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \boldsymbol{\lambda}^T(\mathrm{A}\boldsymbol{x} - \mathbf{b}) \qquad \text{(Unconstrained CVX)}$$

According to first-order optimality,

$$\nabla f(\boldsymbol{x}) + A^T\boldsymbol{\lambda} = 0$$

which admits a solution $\boldsymbol{x}$ satisfies $A\boldsymbol{x} = \boldsymbol{b}$ and

$$\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0 \text{ for all } \boldsymbol{y} \text{ such that } A\boldsymbol{y} = \boldsymbol{b}$$

Lagrangian Multiplier Method

Consider general minimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0,\ i = 1, \ldots, m \\ & \ell_j(x) = 0,\ j = 1, \ldots, r \end{aligned}$$

Need not be convex, but of course we will pay special attention to convex case

We define the Lagrangian as

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

New variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \ge 0$ (else $L(x, u, v) = -\infty$)

Lagrangian Multiplier Method

Important property: for any $u \geq 0$ and v,

$$f(x) \geq L(x, u, v) \quad \text{at each feasible } x$$

Why? For feasible x,

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i \underbrace{h_i(x)}_{\leq 0} + \sum_{j=1}^{r} v_j \underbrace{\ell_j(x)}_{=0} \leq f(x)$$

Let C denote primal feasible set, $f^\star$ denote primal optimal value. Minimizing $L(x, u, v)$ over all x gives a lower bound:

$$f^\star \geq \min_{x \in C} L(x, u, v) \geq \min_{x} L(x, u, v) := g(u, v)$$

Lagrangian Multiplier Method

- Minimize $f(x, y) = x^2 + y^2$ subject to the constraint $x + y = 4$
- **Solution**:

Step 1: Form the Lagrangian function

$L(x, y, \lambda) = f(x, y) - \lambda(g(x, y))$

where $g(x, y)$ is the constraint equation in the form $g(x, y) = 0$

$L(x, y, \lambda) = x^2 + y^2 - \lambda(x + y - 4)$

Step 2: Take partial derivatives and set them equal to zero

$\frac{\partial L}{\partial x} = 2x - \lambda = 0$

$\frac{\partial L}{\partial y} = 2y - \lambda = 0$

$\frac{\partial L}{\partial \lambda} = -(x + y - 4) = 0$

Step 3: Solve the system of equations

From the first equation: $x = \frac{\lambda}{2}$

From the second equation: $y = \frac{\lambda}{2}$

Therefore, $x = y$

Substituting into the third equation:

$x + y = 4$

$2x = 4$

$x = 2$

Since $x = y$, then $y = 2$

And $\lambda = 2x = 4$

Step 4: Verify this is a minimum

The critical point is $(2, 2)$

Let's verify it's a minimum visually with the graph of the constraint and level curves of the objective function

Lagrangian Multiplier Method

- Minimize $f(x, y) = xy$ subject to the constraint $x^2 + y^2 = 16$
- **Solution**:

Step 1: Form the Lagrangian function

$L(x, y, \lambda) = f(x, y) - \lambda(g(x, y))$

$L(x, y, \lambda) = xy - \lambda(x^2 + y^2 - 16)$

Step 2: Take partial derivatives and set equal to zero

$\frac{\partial L}{\partial x} = y - 2\lambda x = 0$ ...(1)

$\frac{\partial L}{\partial y} = x - 2\lambda y = 0$ ...(2)

$\frac{\partial L}{\partial \lambda} = -(x^2 + y^2 - 16) = 0$ ...(3)

Step 3: Solve the system

From (1): $y = 2\lambda x$

From (2): $x = 2\lambda y$

Combining these:

$y = 2\lambda x$ and $x = 2\lambda y$

Therefore: $y = 2\lambda(2\lambda y)$

$y = 4\lambda^2 y$

$1 = 4\lambda^2$ (since $y \neq 0$ for maximum)

$\lambda = \pm\frac{1}{2}$

When $\lambda = \frac{1}{2}$:

$y = 2(\frac{1}{2})x = x$

Substituting into constraint equation (3):

$x^2 + x^2 = 16$

$2x^2 = 16$

$x^2 = 8$

$x = \pm 2\sqrt{2}$

Since $y = x$, we have points $(2\sqrt{2}, 2\sqrt{2})$ and $(-2\sqrt{2}, -2\sqrt{2})$

When $\lambda = -\frac{1}{2}$:

$y = -x$

Leading to points $(2\sqrt{2}, -2\sqrt{2})$ and $(-2\sqrt{2}, 2\sqrt{2})$

Let's evaluate $f(x, y) = xy$ at these points:

At $(2\sqrt{2}, 2\sqrt{2})$ and $(-2\sqrt{2}, -2\sqrt{2})$: $f(x, y) = 8$

At $(2\sqrt{2}, -2\sqrt{2})$ and $(-2\sqrt{2}, 2\sqrt{2})$: $f(x, y) = -8$

Conjugate Functions

- The conjugate of a function f is

$$f^* = \sup_{\mathbf{x} \in dom f} (\mathbf{y}^T \mathbf{x} - f(\mathbf{x}))$$

f^* is convex (even when f is not)

Fenchel's inequality: the definition implies that

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^T \mathbf{y}, \text{for all } \mathbf{x}, \mathbf{y} \in dom f$$

This is an extension to non-quadratic convex f of the inequality

$$\frac{1}{2}\mathbf{x}^T\mathbf{x} + \frac{1}{2}\mathbf{y}^T\mathbf{y} \geq \mathbf{x}^T\mathbf{y}$$

Conjugate Functions

- **Theorem**: The conjugate function $f^*(y)$ is ALWAYS convex of whether the original function $f(x)$ is convex or not.

- *Proof*: The conjugate function is defined as:

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{xy - f(x)\}$$

For each fixed x, the function $xy - f(x)$ is a linear function of y (thus convex in y)

The supremum of any collection of convex functions is always convex. This is because:

For any two points y_1, y_2 and $\lambda \in [0,1]$:

$f^*(\lambda y_1 + (1-\lambda)y_2) = \sup\{x(\lambda y_1 + (1-\lambda)y_2) - f(x)\}$

$= \sup\{\lambda(xy_1 - f(x)) + (1-\lambda)(xy_2 - f(x))\}$

$\leq \lambda\sup\{xy_1 - f(x)\} + (1-\lambda)\sup\{xy_2 - f(x)\}$

$= \lambda f^*(y_1) + (1-\lambda)f^*(y_2)$

Conjugate Function Examples

- Compute the conjugate function of $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$

First, substitute $f(x)$:

$$f^*(y) = \sup_x (y^T x - \frac{1}{2}x^T Ax - b^T x - c)$$

To find the supremum, we take the derivative with respect to x and set it to zero:

$$\frac{\partial}{\partial x}(y^T x - \frac{1}{2}x^T Ax - b^T x - c) = 0$$

$$y - Ax - b = 0$$

Since A is assumed to be symmetric (as it appears in a quadratic form), we can solve for x:

$$x = A^{-1}(y - b)$$

Substitute this x back into the original expression:

$$f^*(y) = y^T[A^{-1}(y-b)] - \frac{1}{2}[A^{-1}(y-b)]^T A[A^{-1}(y-b)] - b^T[A^{-1}(y-b)] - c$$

Simplify:

$$f^*(y) = y^T A^{-1}y - y^T A^{-1}b - \frac{1}{2}(y-b)^T A^{-1}(y-b) - b^T A^{-1}y + b^T A^{-1}b - c$$

Further simplification leads to:

$$\boxed{f^*(y) = \frac{1}{2}(y-b)^T A^{-1}(y-b) - c}$$

This is the conjugate function of the given quadratic function. Note that this assumes A is positive definite (to ensure the supremum exists and A is invertible).

Conjugate Function Examples

- Compute the conjugate function of Negative Entropy $f(x) = \sum_{i=1}^{n} x_i \log x_i$, s.t., $\sum_i x_i = 1$

The conjugate function is defined as:

$$f^*(y) = \sup_x\{\langle y, x\rangle - f(x)\} = \sup_x\{\sum_i y_i x_i - \sum_i x_i \log(x_i)\}$$

subject to constraints: $\sum_i x_i = 1$ and $x_i \geq 0$

Using Lagrangian multipliers:

$$L(x, \lambda) = \sum_i y_i x_i - \sum_i x_i \log(x_i) - \lambda(\sum_i x_i - 1)$$

Taking partial derivatives with respect to x_i and setting to zero:

$$\frac{\partial L}{\partial x_i} = y_i - (\log(x_i) + 1) - \lambda = 0$$

$$\log(x_i) + 1 = y_i - \lambda$$

$$x_i = e^{y_i - \lambda - 1}$$

Using the constraint $\sum_i x_i = 1$:

$$\sum_i e^{y_i - \lambda - 1} = 1$$

$$e^{-\lambda-1} \sum_i e^{y_i} = 1$$

$$e^{-\lambda-1} = \frac{1}{\sum_i e^{y_i}}$$

Therefore:

$$x_i = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

Substituting back into the conjugate function:

$$f^*(y) = \sum_i y_i \frac{e^{y_i}}{\sum_j e^{y_j}} - \sum_i \frac{e^{y_i}}{\sum_j e^{y_j}} \log(\frac{e^{y_i}}{\sum_j e^{y_j}})$$

After simplification:

$$f^*(y) = \log(\sum_i e^{y_i})$$

Therefore, the conjugate function of negative entropy is:

$$\boxed{f^*(y) = \log(\sum_i e^{y_i})}$$

Conjugate Function Examples

- Compute the conjugate function of Negative Logarithm $f(x) = -ln(x)$, defined on $\mathbb{R}_{++}$

The conjugate function $f^*(y)$ is defined as:

$$f^*(y) = \sup_{x>0}\{yx - f(x)\} = \sup_{x>0}\{yx + \ln(x)\}$$

Let's solve this step by step:

To find the supremum, we differentiate with respect to x and set it equal to zero:

$$\frac{d}{dx}(yx + \ln(x)) = y + \frac{1}{x} = 0$$

Solving for x:

$$y + \frac{1}{x} = 0$$
$$\frac{1}{x} = -y$$
$$x = -\frac{1}{y}$$

For this to be a maximum and valid (since x must be positive), we need $y < 0$

For this to be a maximum and valid (since x must be positive), we need $y < 0$

Substituting this value of x back into the original expression:

$$\begin{aligned} f^*(y) &= y(-\frac{1}{y}) + \ln(-\frac{1}{y}) \\ &= -1 + \ln(\frac{-1}{y}) \\ &= -1 - \ln(y) - \ln(-1) \\ &= -1 - \ln(-y) + i\pi \end{aligned}$$

Therefore, the conjugate function is:

$$f^*(y) = \begin{cases} -1 - \ln(-y) & \text{if } y < 0 \\ +\infty & \text{if } y \geq 0 \end{cases}$$

Dual Norm

- **Definition**: For a norm$\|\cdot\|$ on a vector space V, its dual norm$\|\cdot\|_*$ on the dual space V^* is defined as:

$$\|y\|_* = \sup\{\langle y, x\rangle : \|x\| \leq 1\}$$

- **Properties**:
  - For p-norms, the dual norm of l_p is l_q, where $\frac{1}{p} + \frac{1}{q} = 1$
  - The dual norm of l_1 is l_∞ and vice versa
  - The dual norm is always convex, even if the original norm is not
  - Cauchy-Swartz inequality: $\langle u, v\rangle \leq \|u\|_*\|v\|$

This is because if we let $x' = \frac{x}{\|x\|}$ (which has norm 1), then:

$$\langle y, x'\rangle \leq \|y\|_*$$

$$\langle y, \frac{x}{\|x\|}\rangle \leq \|y\|_*$$

$$\langle y, x\rangle \leq \|y\|_*\|x\|$$

Dual Norm

Proposition: The dual norm of l_1 is l_∞

Proof:

For the L1 norm constraint:

$$\|x\|_1 \le 1 \iff \sum_{i=1}^{n} |x_i| \le 1$$

We can rewrite the dual norm as:

$$\|y\|_* = \sup\{\sum_{i=1}^{n} y_i x_i : \sum_{i=1}^{n} |x_i| \le 1\}$$

For any feasible x, we have:

$$|\sum_{i=1}^{n} y_i x_i| \le \sum_{i=1}^{n} |y_i x_i| \le \|y\|_\infty \sum_{i=1}^{n} |x_i| \le \|y\|_\infty$$

where $\|y\|_\infty = \max_i |y_i|$

This shows that $\|y\|_* \le \|y\|_\infty$

Gradient descent

Consider unconstrained, smooth convex optimization

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

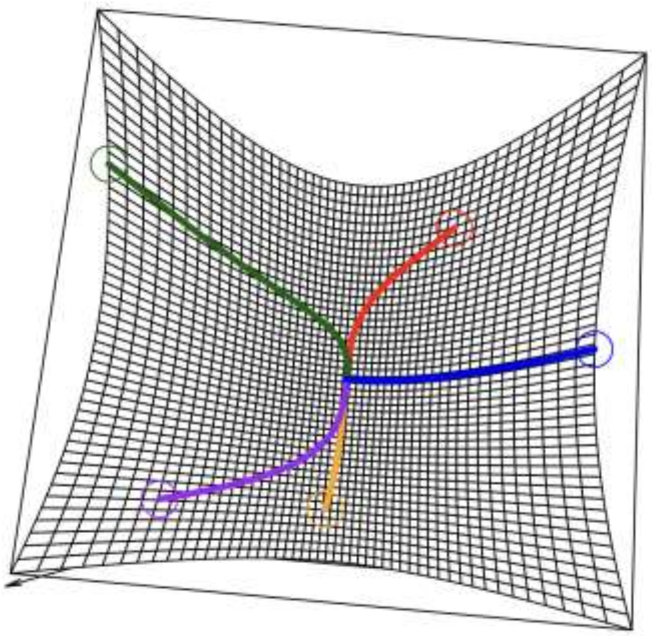
That is, f is convex and differentiable with $\mathrm{dom}(f) = \mathbb{R}^n$. Denote optimal criterion value by $f^\star = \min_x \ f(x)$, and a solution by $x^\star$

Gradient descent: choose initial point $x^{(0)} \in \mathbb{R}^n$, repeat:

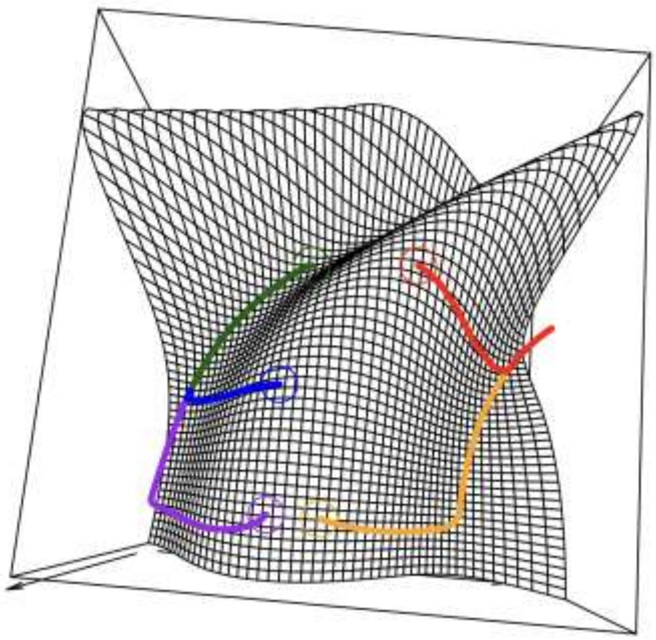
$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Stop at some point

Gradient descent



Convex case



Non-convex case

Gradient descent interpretation

At each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

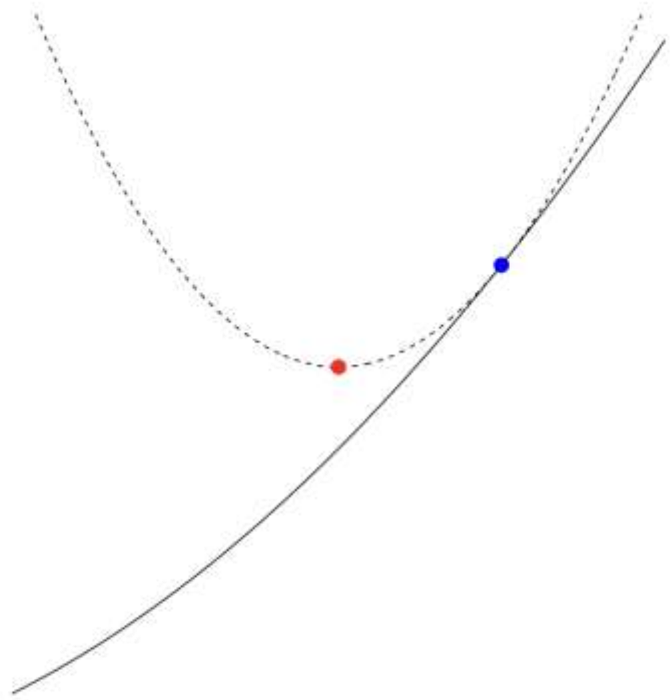
Quadratic approximation, replacing usual Hessian $\nabla^2 f(x)$ by $\frac{1}{t} I$

$f(x) + \nabla f(x)^T (y - x)$	linear approximation to f
$\frac{1}{2t} \|y - x\|_2^2$	proximity term to x, with weight $1/(2t)$

Choose next point $y = x^+$ to minimize quadratic approximation:

$$x^+ = x - t \nabla f(x)$$

Gradient descent interpretation



Blue point is x, red point is

$$x^+ = \underset{y}{\operatorname{argmin}} \; f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

Convergence analysis

Assume that f convex and differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally that ∇f is Lipschitz continuous with constant $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \text{ for any } x, y$$

(Or when twice differentiable: $\nabla^2 f(x) \preceq LI$)

Theorem: Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

and same result holds for backtracking, with t replaced by β/L

We say gradient descent has convergence rate $O(1/k)$. That is, it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations

Convergence analysis

Gradient descent has $O(1/\epsilon)$ convergence rate over problem class of convex, differentiable functions with Lipschitz gradients

First-order method: iterative method, which updates $x^{(k)}$ in

$$x^{(0)} + \operatorname{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots \nabla f(x^{(k-1)})\}$$

Theorem (Nesterov): For any $k \leq (n-1)/2$ and any starting point $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies

$$f(x^{(k)}) - f^\star \geq \frac{3L\|x^{(0)} - x^\star\|_2^2}{32(k+1)^2}$$

Stochastic gradient descent

Consider minimizing an average of functions

$$\min_x \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

As $\nabla \sum_{i=1}^m f_i(x) = \sum_{i=1}^m \nabla f_i(x)$, gradient descent would repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

In comparison, stochastic gradient descent or SGD (or incremental gradient descent) repeats:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f_{i_k}(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

where $i_k \in \{1, \ldots, m\}$ is some chosen index at iteration k

Two rules for choosing index i_k at iteration k:

- Randomized rule: choose $i_k \in \{1, \ldots, m\}$ uniformly at random
- Cyclic rule: choose $i_k = 1, 2, \ldots, m, 1, 2, \ldots, m, \ldots$

Randomized rule is more common in practice. For randomized rule, note that

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$$

so we can view SGD as using an unbiased estimate of the gradient at each step

Main appeal of SGD:

- Iteration cost is independent of m (number of functions)
- Can also be a big savings in terms of memory useage

Example: stochastic logistic regression

Given $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\}$, $i = 1, \ldots, n$, recall logistic regression:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\Big(-y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta)) \Big)}_{f_i(\beta)}$$

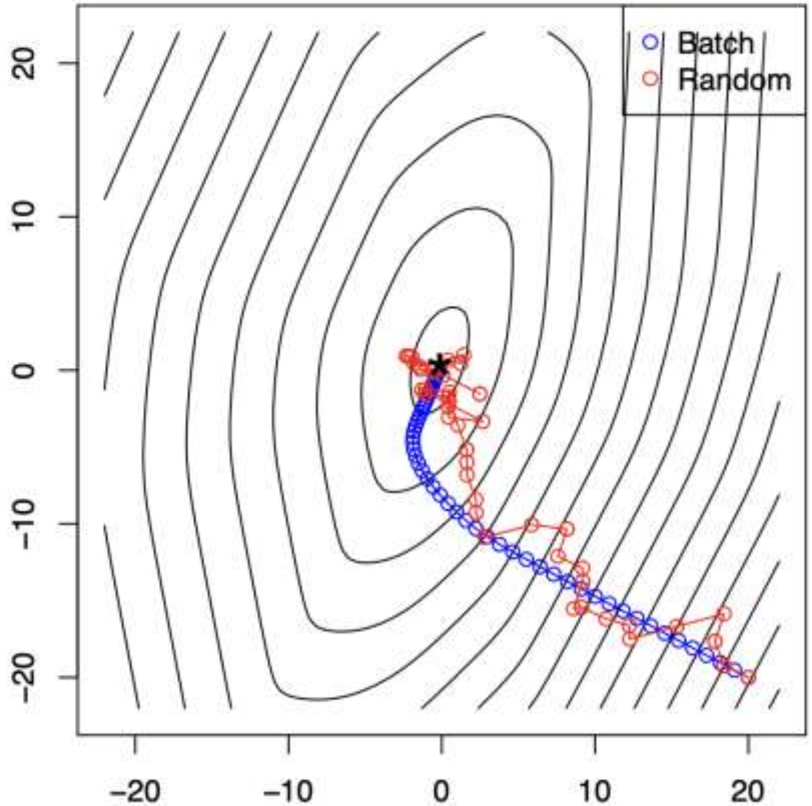
Gradient computation $\nabla f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - p_i(\beta)\big) x_i$ is doable when n is moderate, but not when n is huge

Full gradient (also called batch) versus stochastic gradient:

- One batch update costs $O(np)$
- One stochastic update costs $O(p)$

Clearly, e.g., 10K stochastic steps are much more affordable

Small example with $n = 10$, $p = 2$ to show the "classic picture" for batch versus stochastic methods:



Blue: batch steps, $O(np)$
Red: stochastic steps, $O(p)$

Rule of thumb for stochastic methods:

- generally thrive far from optimum
- generally struggle close to optimum

Convergence rates

Recall: for convex f, gradient descent with diminishing step sizes satisfies

$$f(x^{(k)}) - f^\star = O(1/\sqrt{k})$$

When f is differentiable with Lipschitz gradient, we get for suitable fixed step sizes

$$f(x^{(k)}) - f^\star = O(1/k)$$

What about SGD? For convex f, SGD with diminishing step sizes satisfies[1]

$$\mathbb{E}[f(x^{(k)})] - f^\star = O(1/\sqrt{k})$$

Unfortunately this does not improve when we further assume f has Lipschitz gradient

Mini-batches

Also common is mini-batch stochastic gradient descent, where we choose a random subset $I_k \subseteq \{1, \dots, m\}$, $|I_k| = b \ll m$, repeat:

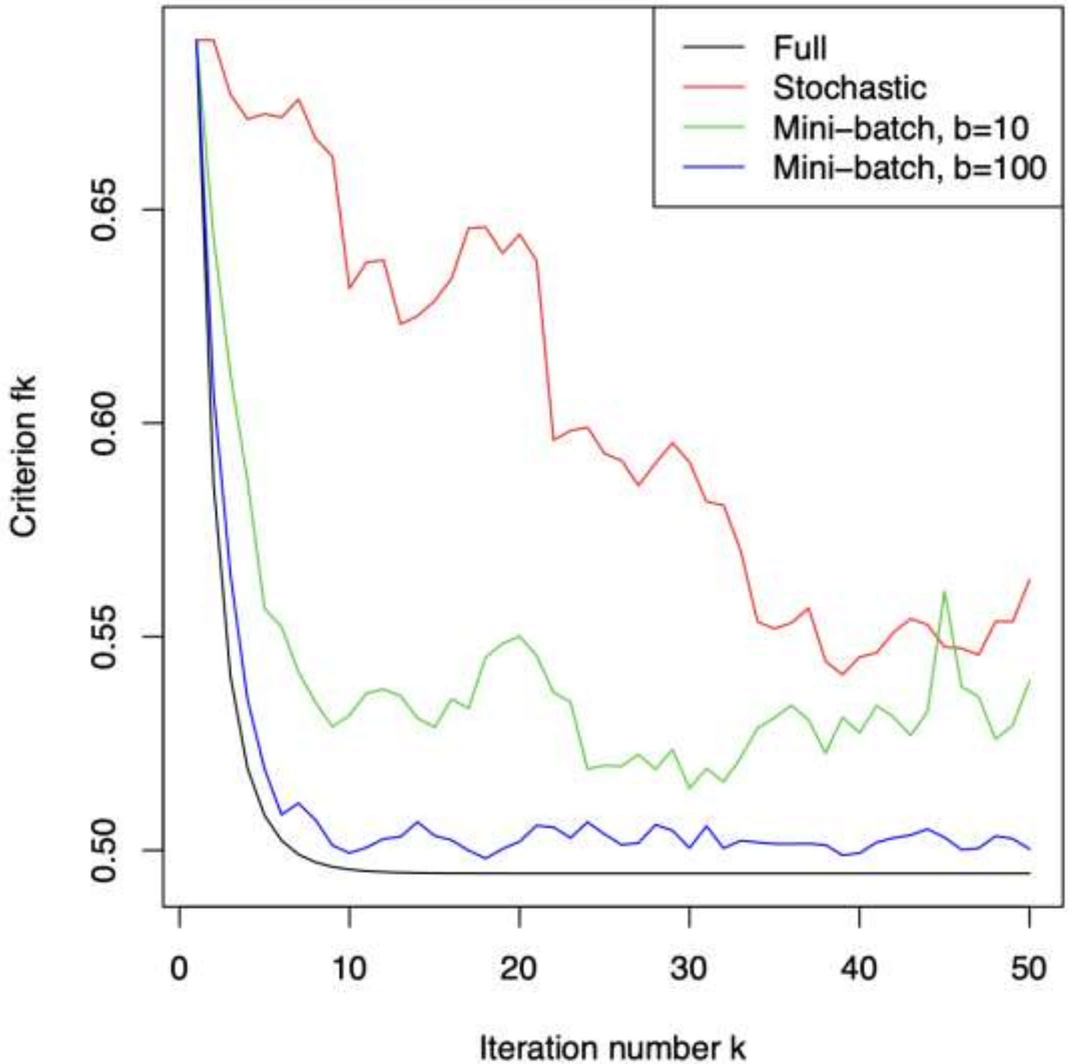
$$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{b} \sum_{i \in I_k} \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \dots$$

Again, we are approximating full gradient by an unbiased estimate:

$$\mathbb{E}\left[\frac{1}{b} \sum_{i \in I_k} \nabla f_i(x)\right] = \nabla f(x)$$

Using mini-batches reduces variance by a factor $1/b$, but is also b times more expensive. Theory is not convincing: under Lipschitz gradient, rate goes from $O(1/\sqrt{k})$ to $O(1/\sqrt{bk} + 1/k)$[3]

Example with $n = 10,000$, $p = 20$, all methods use fixed step sizes:



SGD in large-scale ML

SGD has really taken off in large-scale machine learning

- In many ML problems we don't care about optimizing to high accuracy, it doesn't pay off in terms of statistical performance
- Thus (in contrast to what classic theory says) fixed step sizes are commonly used in ML applications
- One trick is to experiment with step sizes using small fraction of training before running SGD on full data set[4]
- Momentum/acceleration, averaging, adaptive step sizes are all popular variants in practice
- SGD is especially popular in large-scale, continous, nonconvex optimization, but it is still not particular well-understood there (a big open issue is that of implicit regularization)

The End of CVX

- The most essential concepts of CVX:
  - Convex sets, convex functions, 1st and 2nd characterization, Lagrangian Multiplier Method, and Conjugate Functions

- Next week, we will turn to a new Paradigm of Probability and Statistics.