Optimization

Andrew An Bian, Francesco Locatello

23-24 March, 2017

Overview

Convex Sets

Convex Functions

Duality

Descent-based Minimization Methods

Optimization for Matrix Factorization

Convex Sets

A set C is convex if the line segment between any two points in C lies in C

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in C, \forall \lambda \in [0,1] \implies \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in C$$

Convex Sets

A set C is convex if the line segment between any two points in C lies in C

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in C, \forall \lambda \in [0,1] \implies \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in C$$

Convex Combination

$\mathbf{x}$ is a convex combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ if

$$\mathbf{x} = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \ldots + \ldots + \lambda_n \mathbf{x}_n$$

with $\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1, \lambda_i \geq 0$

Some examples

Hyperplanes

A set of the form

$$\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = b\}$$

Some examples

Hyperplanes

A set of the form

$$\left\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = b\right\}$$

Is a hyperplane convex?

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the elements of the hyperplane

$$\begin{aligned}\mathbf{a}^T(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) &= \lambda\mathbf{a}^T\mathbf{x}_1 + (1-\lambda)\mathbf{a}^T\mathbf{x}_2 \\ &= \lambda b + (1-\lambda)b = b\end{aligned}$$

Some examples

Balls

$$B(\mathbf{x}_c, r) = \{\mathbf{x} \;\mid\; \|\mathbf{x} - \mathbf{x}_c\|_2 \leq r\}$$

Some examples

Balls

$$B(\mathbf{x}_c, r) = \{\mathbf{x} \;\mid\; \|\mathbf{x} - \mathbf{x}_c\|_2 \leq r\}$$

Is a ball convex?

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the elements of the ball : $\|\mathbf{x}_i - \mathbf{x}_c\|_2 \leq r$

$$\begin{aligned}\|\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 - \mathbf{x}_c\|_2 &= \|\lambda(\mathbf{x}_1 - \mathbf{x}_c) + (1-\lambda)(\mathbf{x}_2 - \mathbf{x}_c)\|_2 \\ &\leq \lambda\|\mathbf{x}_1 - \mathbf{x}_c\|_2 + (1-\lambda)\|\mathbf{x}_2 - \mathbf{x}_c\|_2 \\ &\leq r\end{aligned}$$

Convexity Preserving Operations

Intersection

If C_s is convex for every $s \in S$ then $\bigcap_{s \in S} C_s$ is convex

Convexity Preserving Operations

Intersection

If C_s is convex for every $s \in S$ then $\bigcap_{s \in S} C_s$ is convex

Affine transformations

If S is convex and $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, $f(S) = \{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ is convex

- Scaling $\alpha S = \{\alpha \mathbf{x} \mid \mathbf{x} \in S\}$
- Translation $S + \mathbf{a} = \{\mathbf{x} + \mathbf{a} \mid \mathbf{x} \in S\}$

Sum of two sets

$$S_1 + S_2 = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in S_1, \mathbf{y} \in S_2\}$$

Non Convex Sets

Union of two sets

Not convex. Let $K = [-1, 0]$ and $L = [1, 2]$. Take one point in K and another in L and draw a line.
You will see that some points on the line fall outside of the union.

Non Convex Sets

Union of two sets

Not convex. Let $K = [-1, 0]$ and $L = [1, 2]$. Take one point in K and another in L and draw a line.
You will see that some points on the line fall outside of the union.

Circle

Take two points on the circle and draw a line joining them.
There are some points on the line which are not on the circle.

Convex Functions

Definition

f is convex if $\forall \mathbf{x}, \mathbf{y}$ and $\lambda \in [0, 1]$

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Epigraph

f is a convex function iff its epigraph is convex

$$\{(\mathbf{x}, t) \mid \mathbf{x} \in domain(f), f(\mathbf{x}) \leq t\}$$

Are these functions convex?

Maximum

$f(x) = \max(x_1, x_2)$

$$\begin{aligned} f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= \max(\lambda_1 x_1 + (1-\lambda)y_1, \lambda x_2 + (1-\lambda)y_2) \\ &\leq \lambda \max(x_1, x_2) + (1-\lambda)\max(y_1, y_2) \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \end{aligned}$$

Are these functions convex?

Maximum

$f(x) = \max(x_1, x_2)$

$$
\begin{aligned}
f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= \max(\lambda_1 x_1 + (1-\lambda)y_1, \lambda x_2 + (1-\lambda)y_2) \\
&\leq \lambda \max(x_1, x_2) + (1-\lambda)\max(y_1, y_2) \\
&= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})
\end{aligned}
$$

$f(u, v) = uv$

Are these functions convex?

Maximum

$f(x) = \max(x_1, x_2)$

$$\begin{aligned} f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= \max(\lambda_1 x_1 + (1-\lambda)y_1, \lambda x_2 + (1-\lambda)y_2) \\ &\leq \lambda \max(x_1, x_2) + (1-\lambda)\max(y_1, y_2) \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \end{aligned}$$

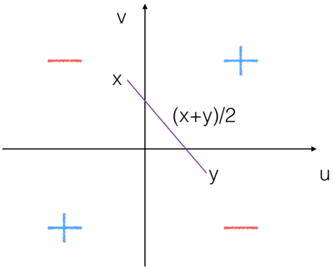
$f(u, v) = uv$

Hint:

- for a convex function we have:

$$f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \leq \frac{f(\mathbf{x}) + f(\mathbf{y})}{2}$$

A Visual Proof



Formal Proof

- To prove convexity, the definition should hold for **ALL** $\mathbf{x}$, $\mathbf{y}$ and λ

Formal Proof

- To prove convexity, the definition should hold for **ALL** $\mathbf{x}$, $\mathbf{y}$ and λ
- To prove non-convexity, it is enough to find **ONE** set of $\mathbf{x}$, $\mathbf{y}$ and λ which does not hold

Formal Proof

- To prove convexity, the definition should hold for **ALL** $\mathbf{x}$, $\mathbf{y}$ and λ
- To prove non-convexity, it is enough to find **ONE** set of $\mathbf{x}$, $\mathbf{y}$ and λ which does not hold
- Take $\lambda = \frac{1}{2}$, $(u, v) = (-1, 1)$, $(u', v') = (1, -1)$

$$0 = f(0,0) \geq \frac{f(-1,1) + f(1,-1)}{2} = -1$$

The Lagrangian

We have an optimization problem (not necessarily convex).

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0 \\ & h_i(\mathbf{x}) = 0 \end{array}$$

Take the constraints into account by augmenting the objective function

Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) + \sum_i \nu_i h_i(\mathbf{x})$$

The Lagrange Dual Function

Dual Function

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- Dual function is always concave!
- It is a lower bound on the optimal value p^* of the original problem for $\boldsymbol{\lambda} \geq 0$:

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$$

Proof of Lower Bound

Suppose $\tilde{\mathbf{x}}$ is a feasible point for the original problem and $\boldsymbol{\lambda} \geq 0$.

- $g_i(\tilde{\mathbf{x}}) \leq 0$ and $h_i(\tilde{\mathbf{x}}) = 0$

Proof of Lower Bound

Suppose $\tilde{\mathbf{x}}$ is a feasible point for the original problem and $\boldsymbol{\lambda} \geq 0$.

- $g_i(\tilde{\mathbf{x}}) \leq 0$ and $h_i(\tilde{\mathbf{x}}) = 0$

$$\sum_i \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_i \nu_i h_i(\tilde{\mathbf{x}}) \leq 0$$

Proof of Lower Bound

Suppose $\tilde{\mathbf{x}}$ is a feasible point for the original problem and $\boldsymbol{\lambda} \geq 0$.

- $g_i(\tilde{\mathbf{x}}) \leq 0$ and $h_i(\tilde{\mathbf{x}}) = 0$

$$\sum_i \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_i \nu_i h_i(\tilde{\mathbf{x}}) \leq 0$$

- $L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\mathbf{x}}) + \sum_i \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_i \nu_i h_i(\tilde{\mathbf{x}}) \leq f(\tilde{\mathbf{x}})$

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\mathbf{x}})$$

Linear Programming Example

We have a Linear Programming problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0, \end{aligned}$$

The Lagrangian is defined by

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T\mathbf{x} + \boldsymbol{\nu}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

Linear Programming Example

We have a Linear Programming problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0, \end{aligned}$$

The Lagrangian is defined by

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T\mathbf{x} + \boldsymbol{\nu}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

The dual function is

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = -\boldsymbol{\nu}^T\mathbf{b} + \inf_{\mathbf{x}} (\mathbf{c}^T - \boldsymbol{\lambda}^T + \boldsymbol{\nu}^T\mathbf{A})\mathbf{x}$$

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\boldsymbol{\nu}^T\mathbf{b} & \text{if } \mathbf{c}^T - \boldsymbol{\lambda}^T + \boldsymbol{\nu}^T\mathbf{A} = 0 \\ -\infty & \text{else} \end{cases}$$

Lagrange Dual Problem

- The dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ gives a lower bound for $\boldsymbol{\lambda} \geq 0$ and $\boldsymbol{\nu}$
- What is the *best* lower bound?

Lagrange Dual Problem

- The dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ gives a lower bound for $\boldsymbol{\lambda} \geq 0$ and $\boldsymbol{\nu}$
- What is the *best* lower bound?

Lagrange Dual Problem

$$\begin{aligned} \underset{\boldsymbol{\lambda}, \boldsymbol{\nu}}{\text{maximize}} \quad & d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0. \end{aligned}$$

Note that this is a convex problem!

Gradient Descent

$$\min_{\mathbf{x} \in \mathbf{R}^d} f(\mathbf{x})$$

- update rule:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \gamma_t \nabla f(\mathbf{x}^t)$$

Gradient Descent: Interpretation

- Remember first order Taylor expansion around $\mathbf{x}$

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

- Now consider parameter $\mathbf{x}^+$ which is obtained by one step of gradient descent.

$$\mathbf{x}^+ = \mathbf{x} - \gamma \nabla f(\mathbf{x})$$

- $\mathbf{x}^+$ minimizes a combination of Taylor expansion penalised by distant to $\mathbf{x}$ as:

$$\mathbf{x}^+ \sim \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})}_{\text{Linear w.r.t } \mathbf{y}} + \frac{1}{\gamma} \underbrace{\|\mathbf{y} - \mathbf{x}\|^2}_{\text{Quadratic}}$$

Gradient Descent: Interpretation

$$\mathbf{x}^+ = \arg\min_{\mathbf{y}} \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})}_{\text{Linear w.r.t } \mathbf{y}} + \frac{1}{\gamma} \underbrace{\|\mathbf{y} - \mathbf{x}\|^2}_{\text{Quadratic}}$$

Question: Using the above interpretation of update rule, explain why choosing smaller step size γ slows down optimization process.

Gradient Descent: Interpretation

$$\mathbf{x}^+ = \arg\min_{\mathbf{y}} \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})}_{\text{Linear w.r.t } \mathbf{y}} + \frac{1}{\gamma} \underbrace{\|\mathbf{y} - \mathbf{x}\|^2}_{\text{Quadratic}}$$

Question: Using the above interpretation of update rule, explain why choosing smaller step size γ slows down optimization process.

Answer: Smaller step size γ leads to a larger weight for the quadratic term that indicates closeness to the current parameter.

Stochastic Gradient Descent

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$$

- Update rule:

$$\mathbf{x}^t = \mathbf{x}^{t-1} - \gamma_t \nabla f_r(\mathbf{x}^t)$$
$$r \text{ uniformly from } \{1, \ldots, n\}$$

Pen & Paper: Unbiased Estimation of Gradient

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

Prove

$$\nabla f(\mathbf{x}) = \mathbf{E}_r\left[\nabla f_r(\mathbf{x})\right]$$

where r is chosen uniformly from $\{1, \ldots, n\}$.

Pen & Paper: Unbiased Estimation of Gradient

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$$

Prove

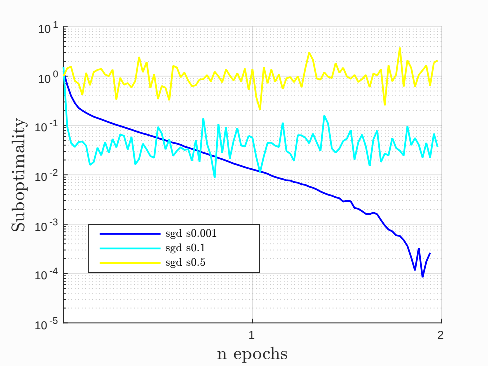
$$\nabla f(\mathbf{x}) = \mathbf{E}_r \left[\nabla f_r(\mathbf{x}) \right]$$

where r is chosen uniformly from $\{1, \ldots, n\}$.
solution:

$$\mathbf{E}_r \left[\nabla f_r(\mathbf{x}) \right] = \sum_i P(r = i) \nabla f_i(\mathbf{x}) = \sum_i \frac{1}{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$$

SGD: step size issue

SGD with a constant step size obtains a suboptimal solution.



SGD: step size issue

SGD with a constant step size obtains a suboptimal solution.

The choice $\gamma_t = 1/t$ guarantees convergence, while it slows down optimization for large t.

Coordinate Descent

$$\min f(x_1, x_2, \ldots, x_d)$$

Update rule:

$$\begin{aligned}
x_1^{t+1} &= \arg\min_{x_1} f(x_1, x_2^t, \ldots, x_d^t) \\
x_2^{t+1} &= \arg\min_{x_2} f(x_1^{t+1}, x_2, x_3^t, \ldots, x_d^t) \\
&\ldots \\
x_d^{t+1} &= \arg\min_{x_d} f(x_1^{t+1}, x_2^{t+1}, \ldots, x_d)
\end{aligned}$$

Pen & Paper: Coordinate Descent

Consider regression problem with the following objective function.

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$$

where $\mathbf{y} \in \mathbf{R}^n$, $\mathbf{A} \in \mathbf{R}^{n \times d}$ with columns $\mathbf{A}_1$, $\mathbf{A}_2$, ..., and $\mathbf{A}_d$.

- Write update rule of coordinate descent

Pen & Paper: Coordinate Descent

Consider regression problem with the following objective function.

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$$

where $\mathbf{y} \in \mathbf{R}^n$, $\mathbf{A} \in \mathbf{R}^{n \times d}$ with columns $\mathbf{A}_1$, $\mathbf{A}_2$, ..., and $\mathbf{A}_d$.

- Write update rule of coordinate descent
- **Solution:**

$$0 \stackrel{!}{=} \nabla_i f(\mathbf{x}) = \mathbf{A}_i^T (\mathbf{A}\mathbf{x} - \mathbf{y}) = \mathbf{A}_i^T (\mathbf{A}_i x_i + \mathbf{A}_{-i}\mathbf{x}_{-i} - \mathbf{y})$$

$$\Leftrightarrow \; x_i = \frac{\mathbf{A}_i^T (\mathbf{y} - \mathbf{A}_{-i}\mathbf{x}_{-i})}{\mathbf{A}_i^T \mathbf{A}_i}$$

Optimization for non-negative matrix factorization

For a given matrix

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|^2$$

$$\text{Subject to } u_{i,j}, v_{j,z} \geq 0$$

where $\mathbf{X}_{n\times m}$,$\mathbf{U}_{n\times k}$ and $\mathbf{V}_{m\times k}$.

- $f(\mathbf{U}, \mathbf{V})$ is convex with respect to $\mathbf{U}$, and convex w.r.t. $\mathbf{V}$. (Why?)
- $f(\mathbf{U}, \mathbf{V})$ is not jointly convex with respect to both of $\mathbf{U}$ and $\mathbf{V}$.
- Alternating minimization is used to optimize the above objective.

Pen & Paper: Collaborative Filtering and SGD

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{|\Omega|} \sum_{(d,n) \in \Omega} \underbrace{\frac{1}{2} \left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{V}^T)_{dn} \right]^2}_{f_{d,n}}$$

and $\mathbf{U} \in Q_1 := \mathbf{R}^{D \times K}$, $\mathbf{V} \in Q_2 := \mathbf{R}^{N \times K}$.

Pen & Paper: Collaborative Filtering and SGD

$$f(\mathbf{U},\mathbf{V}) = \frac{1}{|\Omega|} \sum_{(d,n)\in\Omega} \underbrace{\frac{1}{2}\left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{V}^T)_{dn}\right]^2}_{f_{d,n}}$$

and $\mathbf{U} \in Q_1 := \mathbf{R}^{D\times K}$, $\mathbf{V} \in Q_2 := \mathbf{R}^{N\times K}$.

- **Stochastic Gradient:** For one fixed element (d,n) of the sum, we derive the gradient entry (d',k) of $\mathbf{U}$, that is $\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{V})$, and analogously for the $\mathbf{V}$ part.

$$\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{V}) = \begin{cases} -\left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{V}^T)_{dn}\right] v_{n,k} & \text{if } d' = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial v_{n',k}} f_{d,n}(\mathbf{U},\mathbf{V}) = \begin{cases} -\left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{V}^T)_{dn}\right] u_{d,k} & \text{if } n' = n \\ 0 & \text{otherwise} \end{cases}$$