General

Collapse all

Announcements

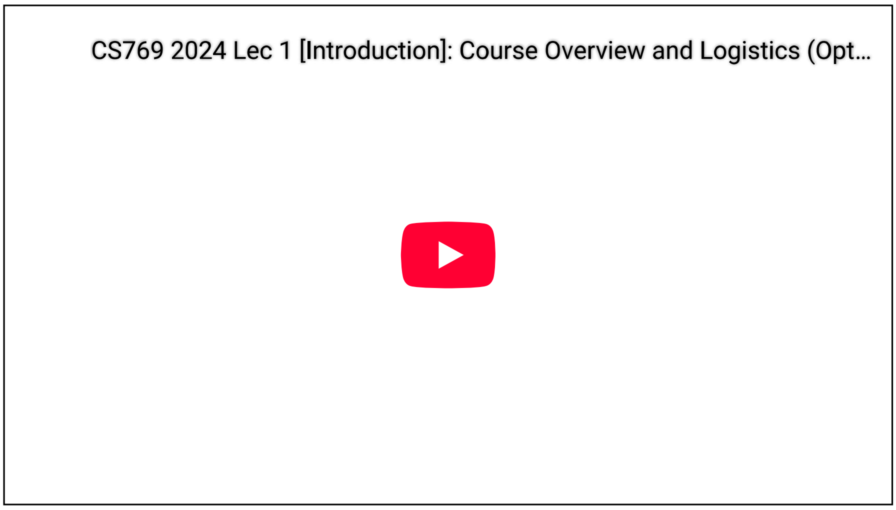
Reference: Playlist of lectures from previous offering of CS709 (CDeep and Youtube)

Mark as done

CS769 Optimization for Machine Learning Playlist on Youtube 2024

Done

Playlist:



;

CS769 Optimization for Machine Learning Playlist on Youtube 2023

Mark as done

Playlist:

CS769 Optimization for Machine Learning Playlist on Youtube 2022

Mark as done

0:00

[Basics of Convex Optimization: My Notes]

Mark as done

[Running Notes: From Basics of ML to Large Language Models [My Notes]]

Mark as done

[Basics of Convex Optimization: My Notes]

Mark as done

[Warmup: ML Book: The Elements of Statistical Learning]

Mark as done

[[Reference Book] Lan G - First-order and stochastic optimization methods for machine learning]

Mark as done

[Convex Optimization Textbook]

Mark as done

[Proximal Sub Gradients, Monotonicity And Convexity]

Mark as done

[Notes on Lipschitz Algebra]

Mark as done

[Sebastian Bubeck's Monograph on Convex Optimization]

Mark as done

[Project slides, report and code submissions]

Mark as done

Opened: Thursday, 1 May 2025, 12:00 AM **Due:** Thursday, 1 May 2025, 12:59 AM

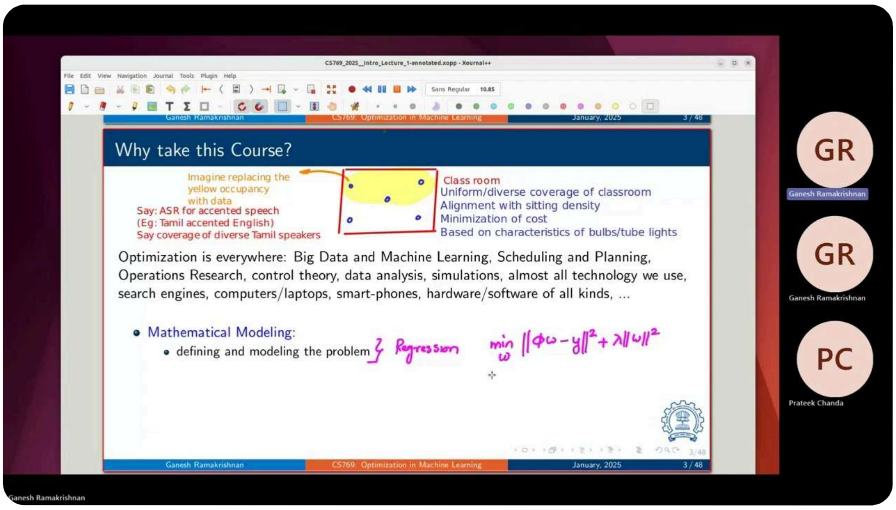
- Project slides, report and code submissions are to be submitted here BEFORE your respective presentations on 2nd May, 3rd May and 5th May as per https://docs.google.com/spreadsheets/d/1gkEhG9jaaneBFI65Geiebn1TUNGheKpjbBQrIQnB8zY/edit?gid=0#gid=0
- Recall the announcement on Friday, 18 April 2025, 1:44 PM that I would like to see all your paper drafts on https://arxiv.org before the presentation toward the submissions at Neurips 2025.
- Additionally, please also upload here before your presentation slot
  1. the link to slides on google drive/pptx and
  2. snapshot pdf of your report here

6 January - 12 January

[Lecture 1: Course Overview, Logistics, ML Supervised Learning Use cases [Monday 6th Jan]]

Mark as done

- Link to video lecture recording from today:

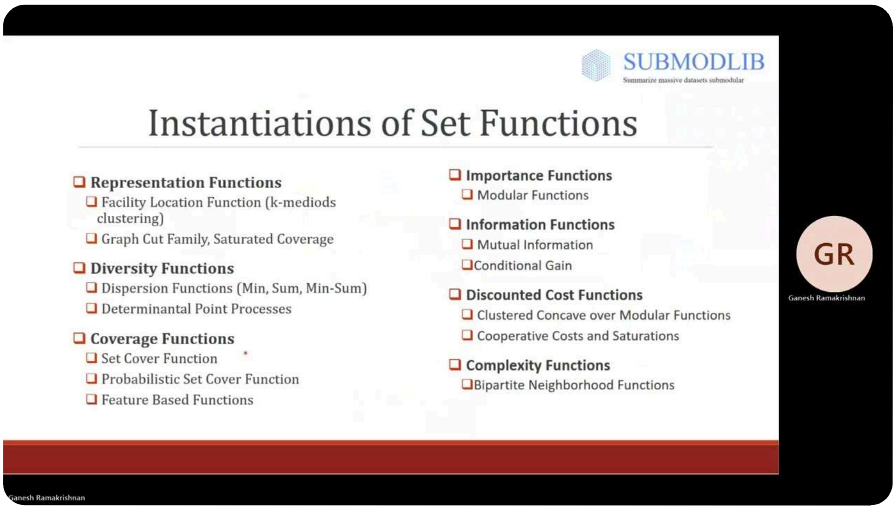


;

- Notes for iteration, editing and discussion by students for Lecture 1 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1xvz4QDCzSeT11n9Y2D-LzyiFiquUs2dfgTtNSZQiGVE/edit?tab=t.0 ;
- CS 769 Course Seminar and Project Themes + Paper List 2025: https://bit.ly/cs769-course-paper-2025 referred to in the slides
- As mentioned today, the lectures will be fast-paced. Students will be referred to previous year lectures as and when required: https://bit.ly/cs769-2024 and https://bit.ly/cs769-2023 and https://bit.ly/cs769-2022

[Previous Year] Lecture on Project ideas: Data Efficient Machine Learning and DECILE https://decile.org/

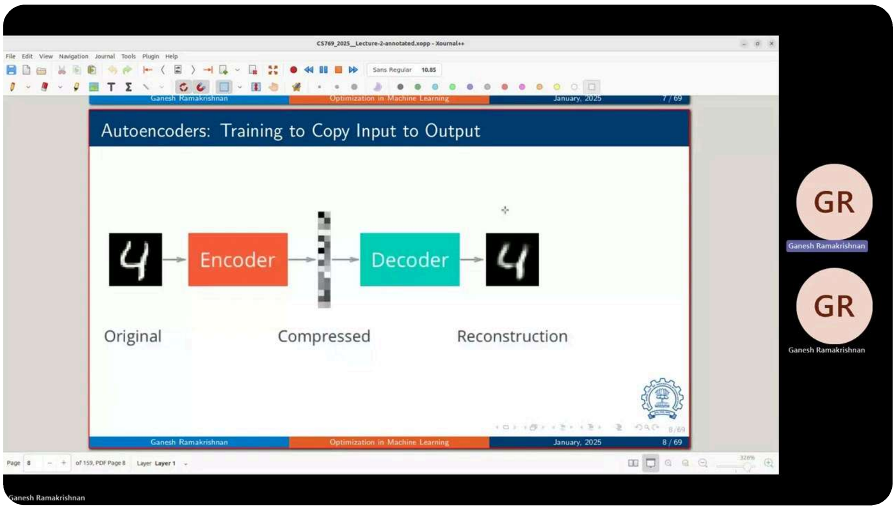
Mark as done



Lecture 2: Continuous Optimization in ML Examples (Part 1) of Supervised Learning, PCA, Matrix Completion, NMF, Clustering

Mark as done

Link to the video lecture:



- Notes for iteration, editing and discussion by students for Lecture 2 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1BG8cD9aokiDgdMYB2R8IXvuqyIZwp9-NTRzvKam8MD8/edit?tab=t.0 ;

Colab Notebooks for brushup

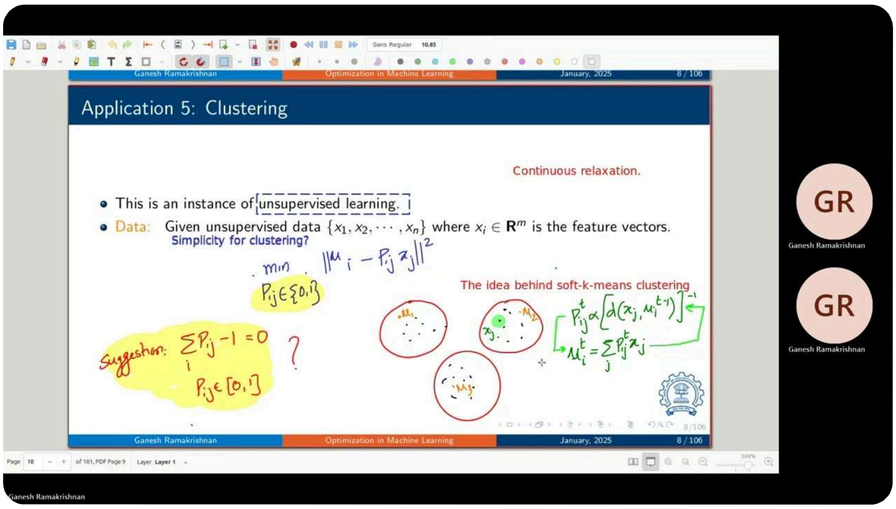
- Numpy-Intro1.ipynb: https://colab.research.google.com/drive/1CdbyGIWVz-iPi_XXVstUt-j_4FFXg3kz#scrollTo=CIPuGOGhTouH
- IntroToPython.ipynb : https://colab.research.google.com/drive/1YU-dL4PrC959-I6J9nOQZAqF8z0-f-Y7#scrollTo=ttg5YC20TH8a
- ML1-LinearRegression.ipynb https://colab.research.google.com/drive/13ILGLUU_k4VS_tCpsMivLc9nnvOluxAN#scrollTo=PyeQbbkjdE6v

13 January - 19 January

PDF **Lecture 3: Continuous Optimization in ML Examples of Clustering and Contextual Bandits (Concluded), Image Segmentation & Correspondence**

Mark as done

Link to video lecture:



;

- Notes for iteration, editing and discussion by students for Lecture 3 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1hhdbu0329ksSMrtlctUC1pVcLadnlenoYNTq_ZGSgzU/edit?tab=t.0
- **Homework: Derive the gradient and the Hessian for the loss functions listed on last page 67 of the slides for Lecture 3.**
- You can refer to Linear Algebra Done Right for brushing up/reading up on linear algebra.
- Calculus brushup for brushing up of principles for optimization of univariate functions
- Linear Algebra Done Right for brushing up of principles for optimization of univariate functions
- *Sample Problem:* Explain the role of the eigenvalues and eigenvectors of the Matrix A in the geometry of the ellipsoid described on page 8 (slide 4) of these notes
- *Sample Problem:* Present a geometric intepretation of the solution to the least squares optimization problem on page 46 (slide 31) of the notes
- *Sample Problem:* Solve exercises on pages 46-46 (slides 27-28): from these notes:

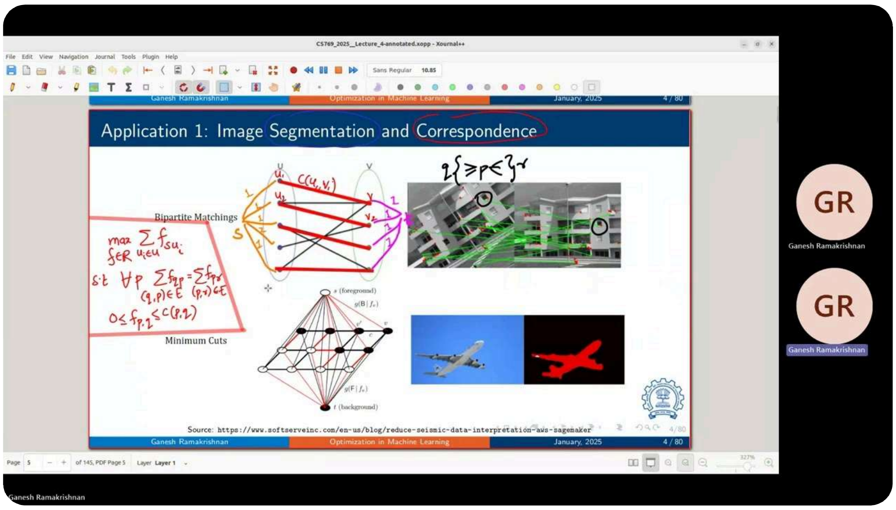
Mark as done

Preparation for Lecture 4

[Lecture 4: Concluding Discrete Optimization Examples (MAP Inference, Feature Selection, Data Subset Selection, etc.), Calculus Brushup for Optimization in Machine Learning, Convex Sets]

Mark as done

Link to video recording:



;

- Notes for iteration, editing and discussion by students for Lecture 4 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1aliWX7fAJRRe3qgLNQJPdY2g4JC9bq9CH7fRRQsYxuw/edit?tab=t.0

Homework:

- **On page 71 of these slides should S be always convex?**
- **On page 71 of these slides, will conv(S) be always convex?**
- **On page 71 of these slides, and based on hint on connection with the separating hyperplane theorm (see page 77) try answering the question: Is there a notion of "Supporting hyperplane" characteristic of convex sets and not found in non-convex sets?**

References (for optional reading)

- Section 4.1.1 (pages 213 to 214) of these notes & Section 4.1.3 (pages 216 to 231) of these notes and you may also watch the video recording here.
- Homework: Derive the gradient and the Hessian for the loss functions listed on page 52 (slide 31) of the slides for Lecture 5.
- You can refer to Linear Algebra Done Right for brushing up/reading up on linear algebra.
- Calculus brushup for brushing up of principles for optimization of univariate functions
- Linear Algebra Done Right for brushing up of principles for optimization of univariate functions
- *Sample Problem:* Explain the role of the eigenvalues and eigenvectors of the Matrix A in the geometry of the ellipsoid described on page 8 (slide 4) of these notes
- *Sample Problem:* Present a geometric intepretation of the solution to the least squares optimization problem on page 46 (slide 31) of the notes
- *Sample Problem:* Solve exercises on pages 46-46 (slides 27-28): from these notes:
- Sections 3.7 until 3.9.4 (pages 175 to 187) of these notes and Sections 3.1 until 3.6 (pages 145 to 175) of these notes.
- Pages 34 to 73 (slides 59 to 98) of these notes.
- Sections 4.2.7 and 4.1.4 of my Convex Optimization notes
- More basic Reference: Refer to Sections 3.1 until 3.6 (pages 145 to 175) of these notes.
- Chapter 2 (From Section 2.1 until 2.2.3) and Chapter 3 (Section 3.1) of Convex Optimization by Boyd and Vandenberghe
- Chapter 3 (Section 3.1) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.7 and 4.1.4 of my Convex Optimization notes

```
Advanced References and Optional Exercises:
```

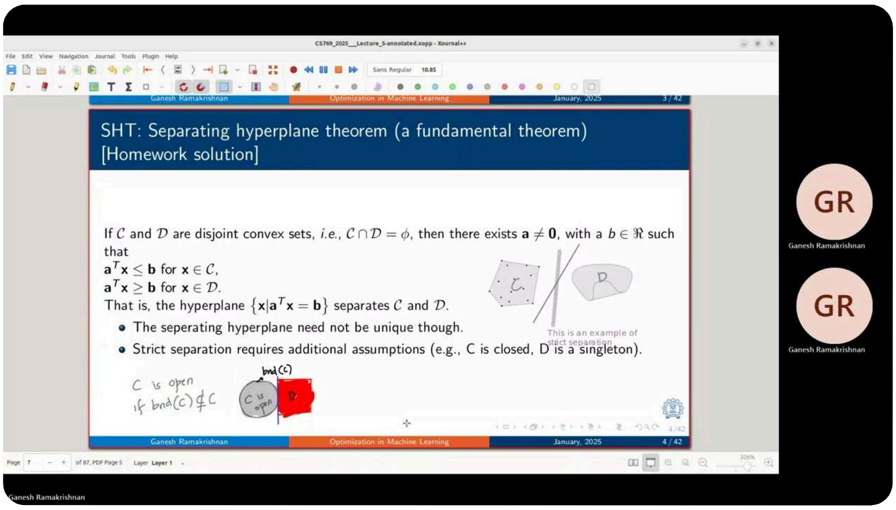
- Advanced Reference on topology, normed spaces, open sets, continuity etc: Geometry I - Basic ideas and concepts of differential geometry by R.V. Gamkrelidze, E. Primrose, D.V. Alekseevskij, V.V. Lychagin, A.M. Vinogradov-
- Advanced Reference on maxima and minima: Stories About Maxima and Minima (Mathematical World) by Vladimir M. Tikhomirov
- Advanced Reference on analysis for optimization: Variational Analysis by Rockafellar, R. Tyrrell, Wets, Roger J.-B.
- Advanced Reference on convex analysis, spaces etc: Fundamentals of Convex Analysis, by Hiriart-Urruty, Jean-Baptiste, Lemarechal, Claude
- Intuitively identify a topological space that is NOT a metric space: referring to page 22 (slide numbered 46 of these notes)
- Intuitively identify a metric space that is NOT a normed space: referring to page 22 (slide numbered 46 of these notes)
- Provide an intuitive explanation why for norms other than the 2-norm, there is no notion of an inner product: referring to page 23 (slide number 47) of these notes

20 January - 26 January

Lecture 5: Convex Functions, Strong Convexity, Calculus of Convexity, ML Examples

Mark as done

Link to Video recording:



;

- Notes for iteration, editing and discussion by students for Lecture 5 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1dWAWV9iLW0TLp6LPEajhXgg0pJnrOx4kN3M1ftfb_Ns/edit?tab=t.0

Homework:

- Determine the convexity (strong/strict etc) of the Machine Learning Objectives on page 31 of the attached slides [https://moodle.iitb.ac.in/pluginfile.php/146795/mod_resource/content/4/CS769_2025__Lecture_5-annotated.pdf]

- Prove that the definition of mid-point convexity (or convexity at 1/2) is equivalent to the general definition of convexity on page 30 of these slides.
- [Optional] Prove that if a function is both concave and convex, it must be affine on page 29 (slide number 179) of these class slides

References for Optional Reading:

- Chapter 3 (Section 3.1) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.7 and 4.1.4 of my Convex Optimization notes

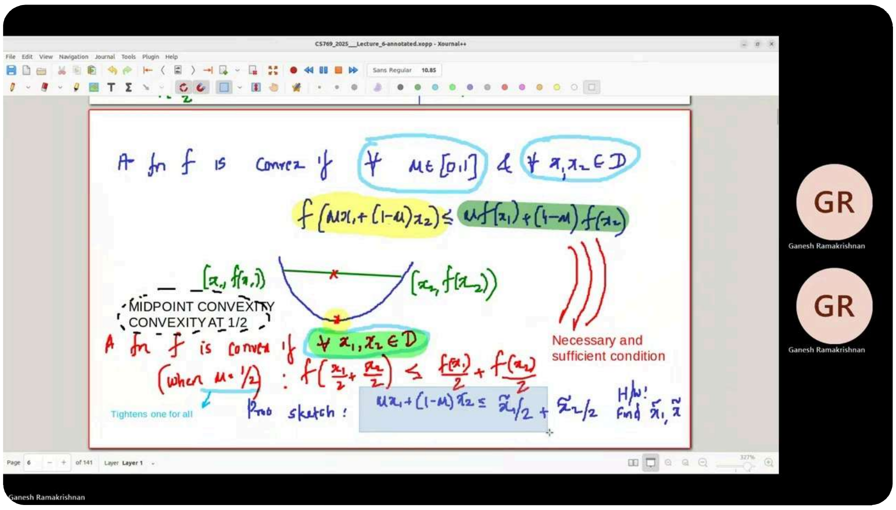
`Advanced References and Optional Exercises:`

- Further reading for proper cones and generalized matrix inequalities: page 34 onwards of these notes and Section 2.6 of Boyd and Vandenberghe
- Optional Problem Solving: Appreciate how the concepts of closed, open sets, interior, boundary etc, extend to general topological/metric spaces.
- Optional Problem Solving: Prove closure of convexity under the perspective function on page 6 (slide number 166) of these class slides

Lecture 6: Calculus of Convexity, ML Examples, Sublevel sets and Epigraphs

Mark as done

Link to video lecture:



Notes for iteration, editing and discussion by students for Lecture 6 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1ftGdazAf975JOPkwyH7aa2cWBdndYyQnkRztYRaoJAo/edit?tab=t.0

Homework:

  - Prove that the negative Gaussian function or its simplified form on pages 44-48 of the slides is quasi-convex. Also is the function is convex?
  - Prove that if a function is both concave and convex, it must be affine on page 29 (slide number 179) of these class slides

References for Optional Reading:

- Chapter 3 (Section 3.1) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.7 and 4.1.4 of my Convex Optimization notes

`Advanced References and Optional Exercises:`

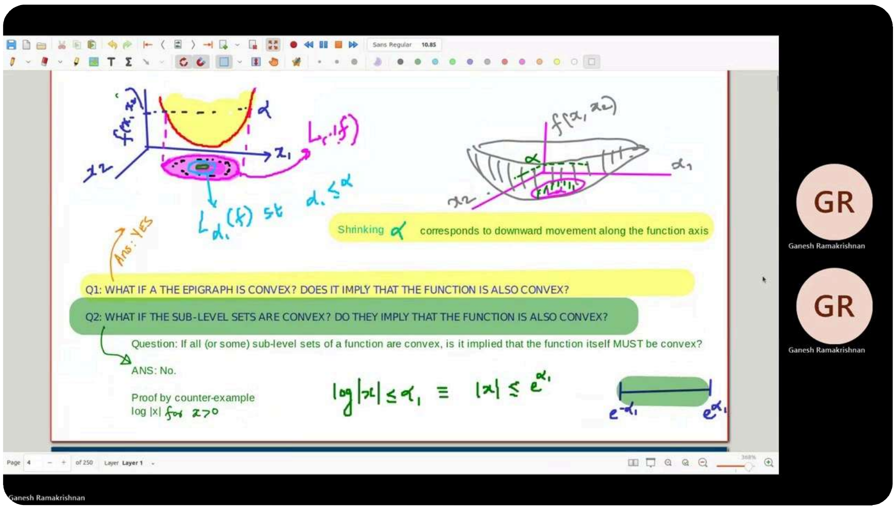
- Further reading for proper conves and generalized matrix inequalities: page 34 onwards of these notes and Section 2.6 of Boyd and Vandenberghe
- Optional Problem Solving: Prove closure of convexity under perspective function on page 6 (slide number 166) of these class slides
- [Optional - for the curious] Relevant lecture from Convex Optimization Course: Lecture 9 is relevant to what is currently running:

27 January - 2 February

Lecture 7: Quasi Convexity, Epigraphs, First and Second order conditions

Mark as done

Link to video lecture:



Notes for iteration, editing and discussion by students for Lecture 7 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/15Yh8pMFGXuRkdD_fBLx9hJQVTr6zQGNCjxGNYYGYO-M/edit?tab=t.0

Homework:

- On page 36, expand the quadratic curvature expression involving the Hessian of logSumExp and ensure that logSumExp can be proved to be convex a
- Make sure you understand the necessary and sufficient conditions along with the proofs for the first order convexity condition on pages 22-34 of the slides.

References (for optional reading)

- Section 4.2.2 of Convex Optimization by Boyd and Vandenberghe
- Sections 3.4, 3.5 (and for further reading, Section 6.5.5) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.8 and 4.1.4 as well as pages 272, 274 of my Convex Optimization notes
- For further reading: Pages 1-8 of notes from S. Boyd, Lecture Notes for EE 264B
- For proof of how the gradient is the unique subgradient at points of differentiabilty of a convex function see pages 4 to 9 of this slide deck

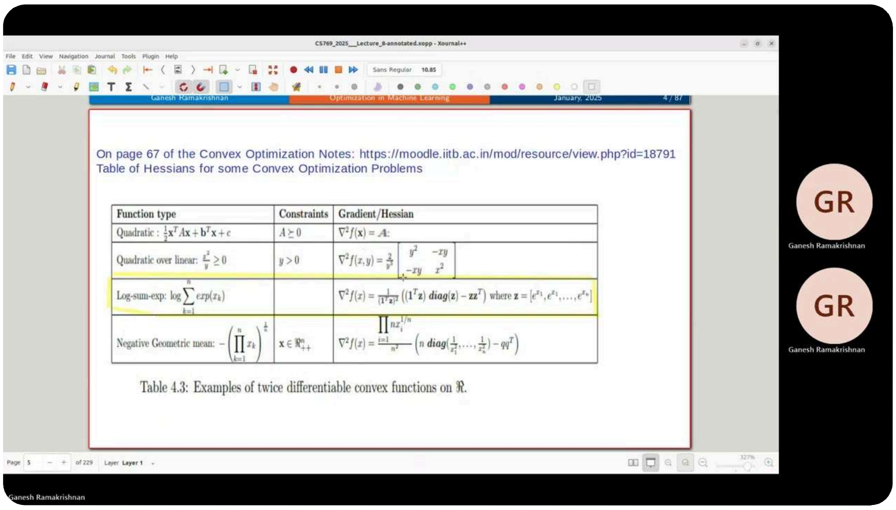
Advanced References and Optional Exercises:

- Optional Problem Solving: Solve the problems (maked as homework) on page 29 (slide number 14) of these class slides
- Optional Problem Solving: Solve the problems (marked as homework) on page 52 (slide number 27) of these class slides
- Understand the simple proof on page 55 (slide number 29) of these class slides
- Optional Problem Solving: Could you think of a non-convex function with non-empty subdifferential a question raised on page 32 (slide number 59) of these class slides
- [Optional - for the curious] Relevant lecture from Convex Optimization Course: Lectures 10 and 11 are relevant to what is currently running:

Lecture 8: Second order Convexity conditions, Subgradient calculus and connections with Gradients, Examples of Subgradients

Mark as done

Link to video lecture :



;

Notes for iteration, editing and discussion by students for Lecture 8 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1-Iv8KRBwfHWhEdyYAaOxNO8FAzWVZP9-YhwFrnCQ2ok/edit?tab=t.0

Homework:

- On page 14, Is the subgradient guaranteed to exist at each point of the domain for a convex function even if the function is non-differentiable?
- Optional: Try and derive the subgradients for the Norm case (page 37) for which, you can use the Holder's inequality stated on the next slide
- Verify the Holder's condition on page 36 of these slides for p=q=2 as well as for p=1, q=infinity

References (for optional reading)

- Section 4.2.2 of Convex Optimization by Boyd and Vandenberghe
- Sections 3.4, 3.5 (and for further reading, Section 6.5.5) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.8 and 4.1.4 as well as pages 272, 274 of my Convex Optimization notes
- For further reading: Pages 1-8 of notes from S. Boyd, Lecture Notes for EE 264B
- For proof of how the gradient is the unique subgradient at points of differentiabilty of a convex function see pages 4 to 9 of this slide deck
- Sections 3.2, 3.4, 3.5 (and for further reading, Section 6.5.5) of Convex Optimization by Boyd and Vandenberghe
- Pages 272, 274, 279-280 in Section 4.2.10 (and also Sections 4.2.8 and 4.1.4) of my Convex Optimization notes
- For further reading: Pages 1-9 of notes from S. Boyd, Lecture Notes for EE 264B

Advanced References and Optional Exercises:

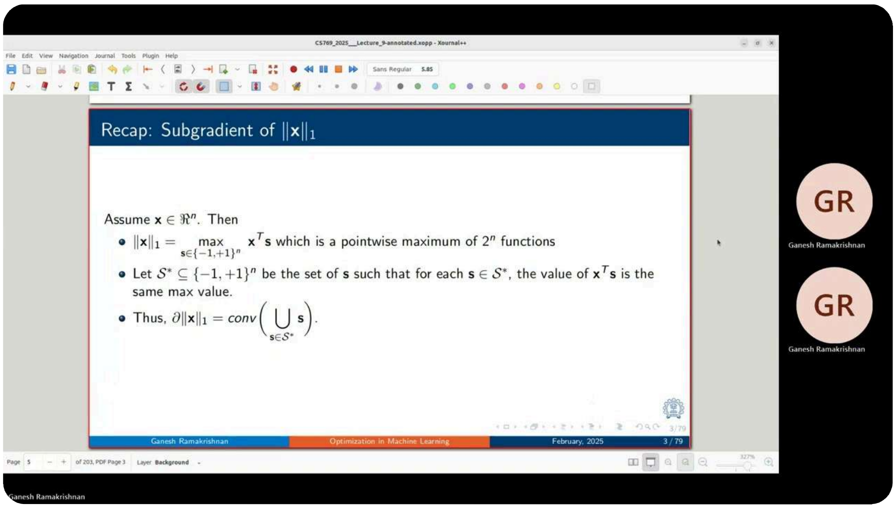
- Optional Problem Solving: Solve the problems (marked as homework) on page 29 (slide number 14) of these class slides
- Optional Problem Solving: Solve the problems (marked as homework) on page 52 (slide number 27) of these class slides
- Understand the simple proof on page 55 (slide number 29) of these class slides
- Optional Problem Solving: Could you think of a non-convex function with non-empty subdifferential a question raised on page 32 (slide number 59) of these class slides
- [Optional - for the curious] Relevant lecture from Convex Optimization Course: Lectures 10 and 11 are relevant to what is currently running:

3 February - 9 February

Lecture 9: Subgradient calculus concluded: Max/Supremum, Composition, Lasso illustration, Further subgradient properties

Mark as done

Link to video lecture:



;

- **Notes for iteration, editing, and discussion by students for Lecture 9 [Students encouraged to add thoughts/including equations or doubts/questions]:** https://docs.google.com/document/d/1sJRjGQaIDtUq7s7SVNOY5GC9qeXhNzjc0-0J9Wy7j70/edit?tab=t.0 ;

Homework:

- Try and Derive the Subgradients on page 21 of the notes.

- Optional: Try and derive the subgradients for the Norm case (page 7) for which, you can use the Holder's inequality stated on the next slide
- Verify the Holder's condition on page 4 of these slides for p=q=2 as well as for p=1, q=infinity
- Please ensure that you understand the subgradient calculus of the proximal operator (pages 23-24)
- Optional: Try and understand advanced & optional properties of the subgradients on pages 25 until 39

References (for optional reading)

- Section 4.2.2 of Convex Optimization by Boyd and Vandenberghe
- Sections 3.4, 3.5 (and for further reading, Section 6.5.5) of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.8 and 4.1.4 as well as pages 272, 274 of my Convex Optimization notes
- For further reading: Pages 1-8 of notes from S. Boyd, Lecture Notes for EE 264B
- For proof of how the gradient is the unique subgradient at points of differentiabilty of a convex function see pages 4 to 9 of this slide deck
- Sections 3.2, 3.4, 3.5 (and for further reading, Section 6.5.5) of Convex Optimization by Boyd and Vandenberghe
- Pages 272, 274, 279-280 in Section 4.2.10 (and also Sections 4.2.8 and 4.1.4) of my Convex Optimization notes
- For further reading: Pages 1-9 of notes from S. Boyd, Lecture Notes for EE 264B

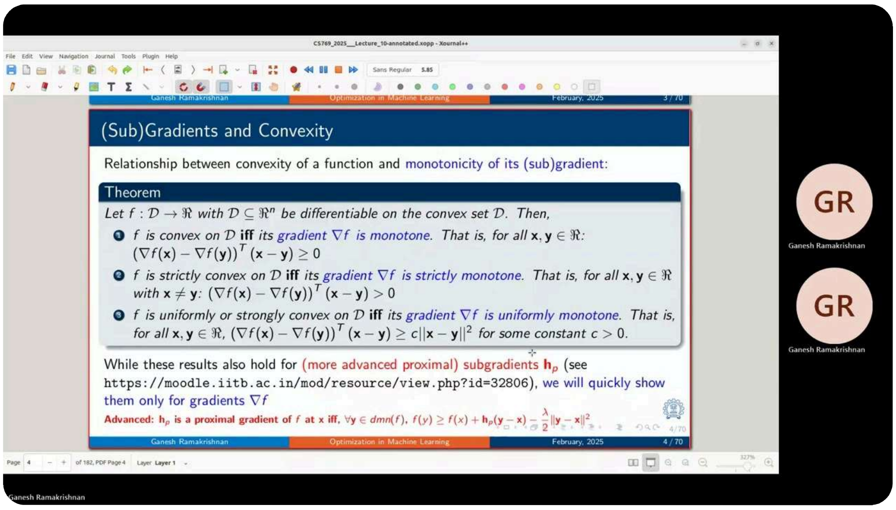
Advanced References and Optional Exercises:

- Optional Problem Solving: Solve the problems (marked as homework) on page 29 (slide number 14) of these class slides
- Optional Problem Solving: Solve the problems (marked as homework) on page 52 (slide number 27) of these class slides
- Understand the simple proof on page 55 (slide number 29) of these class slides
- Optional Problem Solving: Could you think of a non-convex function with non-empty subdifferential a question raised on page 32 (slide number 59) of these class slides
- [Optional - for the curious] Relevant lecture from Convex Optimization Course: Lectures 10 and 11 are relevant to what is currently running:

Lecture 10: Sufficient and Necessary conditions for optimization with and without Convexity

Mark as done

Link to video lecture:



Notes for iteration, editing, and discussion by students for Lecture 10 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1FuhHldaZQmUYz4wXZt3Cbw4Qu6eZNRcvDmbb3yJPGeA/edit?tab=t.0

Homework:

- Think of/enumerate lots of machine objective functions for minimization and try answering the following for each (i) When does global minimum exist (see page 29) (ii) What does Weierstass theorem mean for each function (see page 40)?
- For the claim that for any convex function, any point of local minimum is also a point of its global minimum, we presented a graphically motivated proof pages 26-36. Can you think of some other proof methodology for the same claim for a strictly or strongly convex function.

References (for optional reading)

- Section 4.2.2 of Convex Optimization by Boyd and Vandenberghe
- Sections 4.2.8 and 4.1.4 of my Convex Optimization notes

Advanced References and Optional Exercises:

- Optionally Solve the problems on [page 29 (slide number 14) of these class slides]
- Optionally Solve the problems on [page 52 (slide number 27) of these class slides]
- [Optional - for the curious] Relevant lecture from Convex Optimization Course: Lecture 9 is relevant to what is currently running:

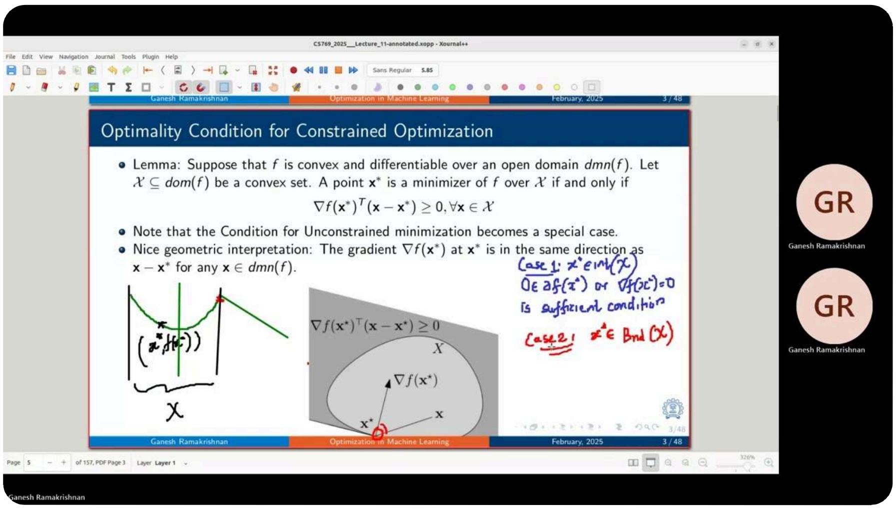
---

10 February - 16 February

Lecture 11: Lipschitz Continuity and Smoothness in Details, Calculus and Examples

Mark as done

Link to video lecture:



- **Notes for iteration, editing, and discussion by students for Lecture 11 [Students encouraged to add thoughts/including equations or doubts/questions]:** [**https://docs.google.com/document/d/1kg1U6jJrY_kq-HPTGCkboNM2OAiZSgmlBBuSCR79btE/edit?tab=t.0**]

Homework:

- Revisit the ML losses discussed and understand if and why they are Lipschitz continuity/smoothness on pages 57-75. Rationalize your answer(s)
- Determine the Lipschitz continuity/smoothness of $$x^4$$ on pages 32-35

References (for optional reading):

- Exhaustive notes on [Lipshictz Algebra]
- [Nice Blog on Lipschitz Continuity]. The author has a [similar blog on Strong Convexity]

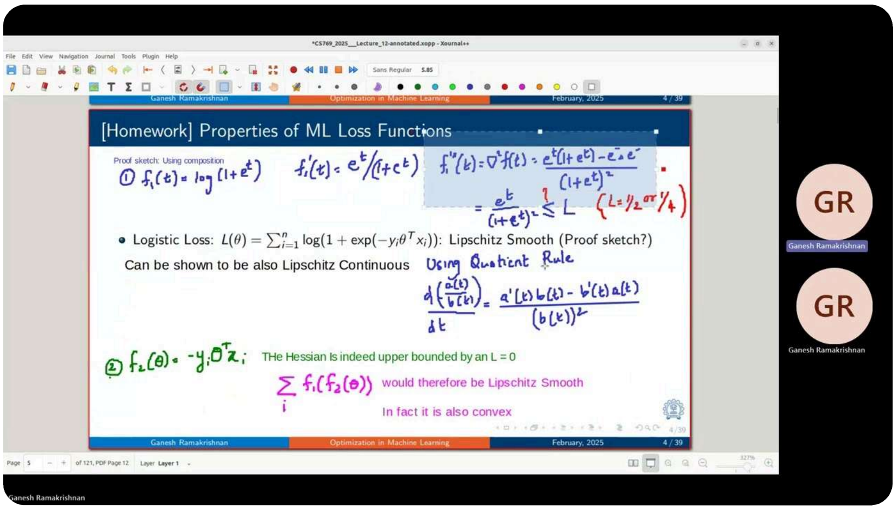
Advanced References and Optional Exercises:

- **Juha Heinonen's, [Lectures on Lipschitz Analysis]**
- **[Anatoli Iouditski's notes]**
- **Optional Problem Solving: What is the connection between Lipschitz continuity of a function and Lipschitz contunity of its gradient. Does one imply the other?**
- **Optional Problem Solving: Show that f is continuous and differentiable everywhere while not being Lipschitz continuous [on page number 31 (slides 136) of these notes]**
- **Optional Problem Solving: Study the plots of functions [on page numbers 26-34 of these notes] for Lipschitz continuity of the function OR its gradient, etc**

Lecture 12: Concluding Lipschitz Continuity and Smoothness illustration on ML objectives, Algorithms for Optimization, Convergence Analysis of Gradient Descent under Lipschitz Continuity and Convexity

Mark as done

Link to video lecture:



- **Notes for iteration, editing, and discussion by students for Lecture 12 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1VYImRgxg110tHD5hhNT-T_V_tGzvJlp9twV8KEu-0zE/edit?tab=t.0**

Homework:

- As per the introspection on page 57, is the analysis of convergence connected more to big-O or Omega?
- Understand how using Lipschitz smoothness, we could faster convegence as hinted on page 33-of this slide deck
- What would be a bound on the value of Lipschitz smootheness constant L for gradient of f when f is the Logistic Loss (pages 4-5)?
- Simpler [already discussed in class]: Under the Analysis, Part III discussed on page 49-52 of this deck, prove that the absolute value of the difference of function at lowest iterate so far and the function at the global minimum satisfies the same upper bound
- Simpler [already discussed in class]: Verify that the expression \frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma} } (named E_2(\gamma)) is minimized with respect to \gamma, by setting its derivative (wrt \gamma) to 0 which is obtained by setting \gamma = \frac{R}{B\sqrt{T}} on pages 49-50 of this deck

[Link to our running Colab Notebook: CS769_sms_data_learning_from_GradientDescent_to_SubmodularOpt.ipynb]()

References (for optional reading):

- **Exhaustive notes on [Lipshictz Algebra]()**
- **[Nice Blog on Lipschitz Continuity](). The author has a [similar blog on Strong Convexity]()**
- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).
- Several graphical illustrations can be found in the matlab code available [here]().
- An illustration of Random walk based descent based on the matlab code available [here](), which is adopted from [matlab code]() accompanying the book [Applied Optimization with MATLAB Programming]()

Advanced References and Optional Exercises:

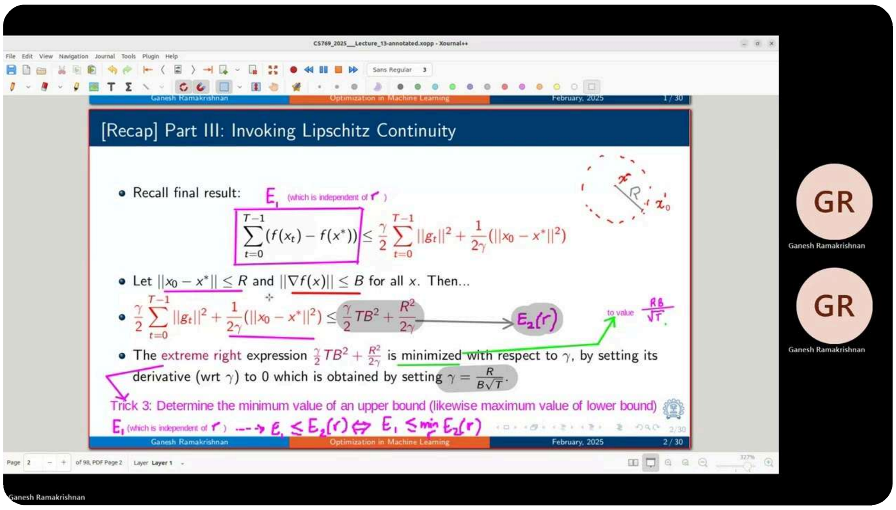
- Juha Heinonen's, [Lectures on Lipschitz Analysis]()
- [Anatoli Iouditski's notes]()
- Optional Problem Solving: What is the connection between Lipschitz continuity of a function and Lipschitz contunity of its gradient. Does one imply the other?
- Optional Problem Solving: Show that f is continuous and differentiable everywhere while not being Lipschitz continuous [on page number 31 (slides 136) of these notes]()
- Optional Problem Solving: Study the plots of functions [on page numbers 26-34 of these notes]() for Lipschitz continuity of the function OR its gradient, etc
- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).
- Several graphical illustrations can be found in the matlab code available [here]().
- An illustration of Random walk based descent based on the matlab code available [here](), which is adopted from [matlab code]() accompanying the book [Applied Optimization with MATLAB Programming]()
- **[See these slides for]()**
  - **Generalized Descent Algorithm and its convergence**
  - **Back-tracking (and exact) ray search, Armijo conditions, Goldstein conditions, etc**
  - **Q-convergence and R-convengence**
- **Nemirovski Section 5.4.2**
- **Optional Problem Solving: Understand and complete the derivations [on page numbers 23-24 of these notes]() for the rate of convergence of the gradient descent algorihtm for convex functions etc**
- **Optional Problem Solving: Constrast the order and rate of convergence of the sequence s2 [on page number 31 (slide 161) of these notes]() against the sequences s1 and r1 discussed a few slides before**
- **Optional Problem Solving: Understand the derivation of the second order condition for convexity [on page numbers 35 to 41 (slides 166-170) of these notes]()**

17 February - 23 February

Lecture 13: Algorithms for Optimization, Enhancements of Convergence Analysis of Gradient Descent via Smoothness and Strong Convexity

Mark as done

Link to video lecture:



Notes for iteration, editing, and discussion by students for Lecture 13 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1WJR-ygjfy79h9vIK9vQGWFpvBzclBwVnnVgRy0aFolw/edit?tab=t.0

Homework:

- **Most important (see page 59): How do we use strong convexity in conjunction with L-smothness to get a sufficient condition as**
  - T >= log(1/ε)
  - Can it be through some intermediate steps culminating in f(μ/L)^T <= ε?
- We saw how how using Lipschitz smoothness, we get faster convegence on pages 13-54 of this slide deck
- Please try and (re) derive bound on the value of Lipschitz smootheness constant L for gradient of f when f is the Logistic Loss (pages 16-18)

Link to our running Colab Notebook: CS769_sms_data_learning_from_GradientDescent_to_SubmodularOpt.ipynb

References (for optional reading):

- **Exhaustive notes on Lipshictz Algebra**
- **Nice Blog on Lipschitz Continuity. The author has a similar blog on Strong Convexity**
- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).
- Several graphical illustrations can be found in the matlab code available here.
- An illustration of Random walk based descent based on the matlab code available here, which is adopted from matlab code accompanying the book Applied Optimization with MATLAB Programming

Advanced References and Optional Exercises:

- Juha Heinonen's, Lectures on Lipschitz Analysis
- Anatoli Iouditski's notes
- Optional Problem Solving: What is the connection between Lipschitz continuity of a function and Lipschitz contunity of its gradient. Does one imply the other?
- Optional Problem Solving: Show that f is continuous and differentiable everywhere while not being Lipschitz continuous on page number 31 (slides 136) of these notes
- Optional Problem Solving: Study the plots of functions on page numbers 26-34 of these notes for Lipschitz continuity of the function OR its gradient, etc
- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).

- Several graphical illustrations can be found in the matlab code available here.
- An illustration of Random walk based descent based on the matlab code available here, which is adopted from matlab code accompanying the book Applied Optimization with MATLAB Programming
- **See these slides for**
  - **Generalized Descent Algorithm and its convergence**
  - **Back-tracking (and exact) ray search, Armijo conditions, Goldstein conditions, etc**
  - **Q-convergence and R-convengence**
- **Nemirovski Section 5.4.2**
- **Optional Problem Solving: Understand and complete the derivations on page numbers 23-24 of these notes for the rate of convergence of the gradient descent algorihtm for convex functions etc**
- **Optional Problem Solving: Constrast the order and rate of convergence of the sequence s2 on page number 31 (slide 161) of these notes against the sequences s1 and r1 discussed a few slides before**
- **Optional Problem Solving: Understand the derivation of the second order condition for convexity on page numbers 35 to 41 (slides 166-170) of these notes**

Lecture 14: Doubts Session + Homework on Convergence Analysis of Gradient Descent Enhancements via Smoothness and Strong Convexity

Mark as done

Link to video lecture:

Notes for iteration, editing, and discussion by students for Lecture 14 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1p-ezJldzD6EwVgC4Eg4eUZDThz9pVyaqgf2RMZwP5Rk/edit?tab=t.0

References (for optional reading):

CS769_sms_RunningNotebook_from_GradientDescent And Demonstration of Different Line Search Appoaches

- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).
- Sections 3.3, 9.3 and 9.4 of Boyd and Vandenberghe and Section 3.4 of Nocedal and Wright
- Sections 1.2.2, 1.2.3, 2.1.3 and 3.2.3 of Introductory Lectures on Convex Programming by Nesterov
- Best convergence rate (lower) bounds for gradient (steepest) descent in section 3.5, page 53 of these notes. Also, bounds with constant step-size from these notes
- For subgradient methods, also refer to Section 1, 2 and 3 of these notes on advanced optimization by Boyd.

Advanced References and Optional Exercises:

- Nemirovski Section 5.4.2
- Optional Problem Solving: Understand the implications of strong convexity on convergence on on page numbers 35-36 (sldes 179-180) of these notes:
- Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
- Optional Problem Solving: Understand the general implications of strong convexity on convergence on on page numbers 35-36 (sldes 179-180) of these notes
- Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
- Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.

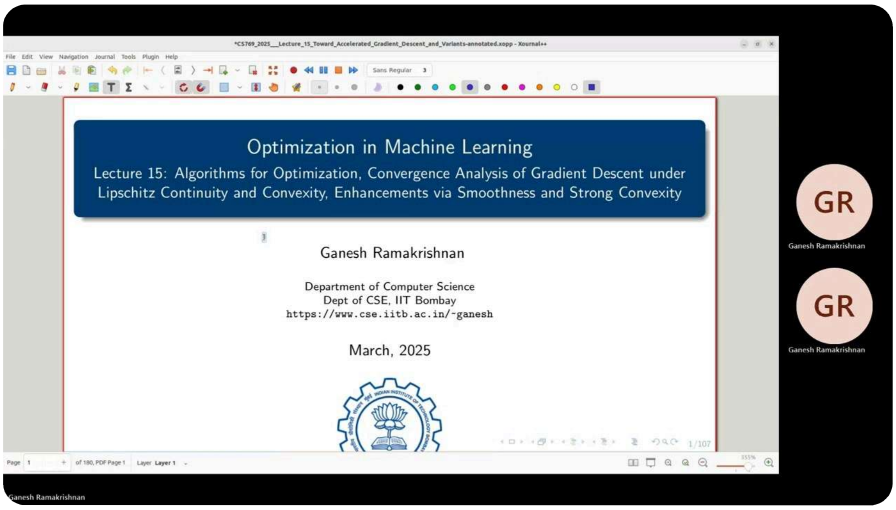
24 February - 2 March

3 March - 9 March

PDF [Lecture 15: Optimization Algorithms and their convergence: Gradient Descent under Strong Convexity + Lipschitz Smoothness/Continuity, Lower Bounds results with convexity, Inspiration for Accelerated GD Variants]

Mark as done

Link to video:



Notes for iteration, editing, and discussion by students for Lecture 15 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1O6mAUu8evv74Lr8b3e2qfoD5nlcNcrjKBrnhjeSriVA/edit?tab=t.0

Homework:

- **Write down the steps for implementing the Nesterov momentum update**
- **How do we reuse different components of analysis of convergence and show the corresponding convergence rate for strong convexity + smoothness starting with page 42 of this deck**
- **Food for thought: What are the fine differences between the Polyak's Heavy ball momentum method and Nesterov's momentum method? Can the gap in the case of smooth function between IDEAL and achieved be bridged using any of these?**

References (for optional reading):

CS769_sms_RunningNotebook_from_GradientDescent And Demonstration of Different Line Search Appoaches

- Boyd and Vandenberghe Sections 9.3.1 and Nocedal and Wright Section A.2 (subsection "Rates of Convergence").
- Nemirovski Section C.4 (appendix).
- Sections 3.3, 9.3 and 9.4 of Boyd and Vandenberghe and Section 3.4 of Nocedal and Wright
- Sections 1.2.2, 1.2.3, 2.1.3 and 3.2.3 of Introductory Lectures on Convex Programming by Nesterov
- Best convergence rate (lower) bounds for gradient (steepest) descent in section 3.5, page 53 of these notes. Also, bounds with constant step-size from these notes
- For subgradient methods, also refer to Section 1, 2 and 3 of these notes on advanced optimization by Boyd.

Advanced References and Optional Exercises:

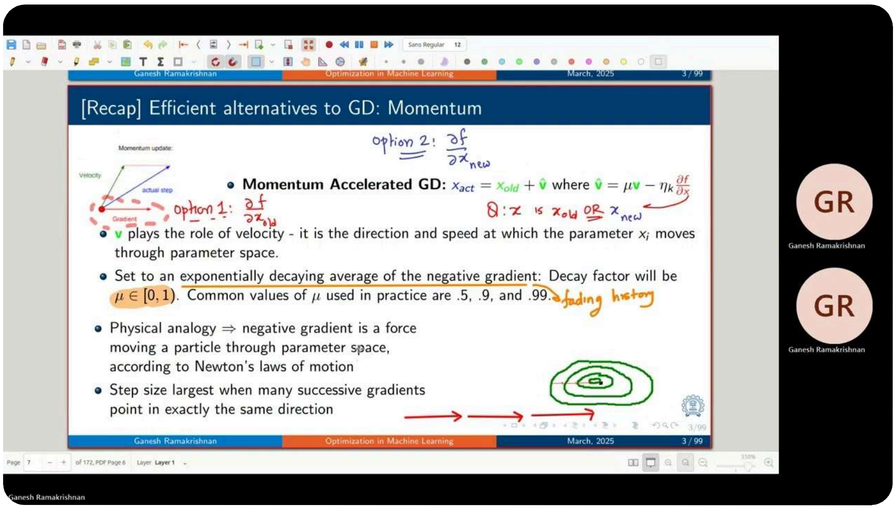
- Nemirovski Section 5.4.2
- Optional Problem Solving: Understand the implications of strong convexity on convergence on on page numbers 35-36 (sldes 179-180) of these notes:
- Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
- Optional Problem Solving: Understand the general implications of strong convexity on convergence on on page numbers 35-36 (sldes 179-180) of these notes
- Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
- Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.

10 March - 16 March

PDF **Lecture 16: Subgradient, Accelerated and Stochastic Gradient Descent**

Mark as done

Video lecture:



Notes for iteration, editing, and discussion by students for Lecture 16 [Students encouraged to add thoughts/including equations or doubts/questions]:

https://docs.google.com/document/d/1_TUDkfjgcfCQbIkjMiZmo8VFO_EbONPDVx8jbzSSfsg/edit?tab=t.0

Homework

- **Understand the convergence analysis of subgradient descent between pages 35 and 49 of these notes/slides**
- **Do all the points matter equally for update. Can we select the points on which updates must be computed. This leads to stochastic GD. Can we analyze stochastic GD for its convergence as on pages 51-62 of these notes/slides? Think about the implications of stochasticity on adaptive methods.**
- **Food for thought: What are the fine differences between the Polyak's Heavy ball momentum method and Nesterov's momentum method? Can the gap in the case of smooth function between IDEAL and achieved be bridged using Polyak?**

References (for optional reading):

- Sections 3.3, 9.3 and 9.4 of Boyd and Vandenberghe and Section 3.4 of Nocedal and Wright
- Sections 1.2.2, 1.2.3, 2.1.3 and 3.2.3 of Introductory Lectures on Convex Programming by Nesterov
- Best convergence rate (lower) bounds for gradient (steepest) descent in section 3.5, page 53 of these notes. Also, bounds with constant step-size from these notes
- For subgradient methods, also refer to Section 1, 2 and 3 of these notes on advanced optimization by Boyd.
- An overview of gradient descent optimization algorithms
- For subgradient methods, also refer to Section 1, 2 and 3 of these notes on advanced optimization by Boyd.
- Chapter on Stochastic Gradient Descent
- Blog on Stochastic Gradient Descent

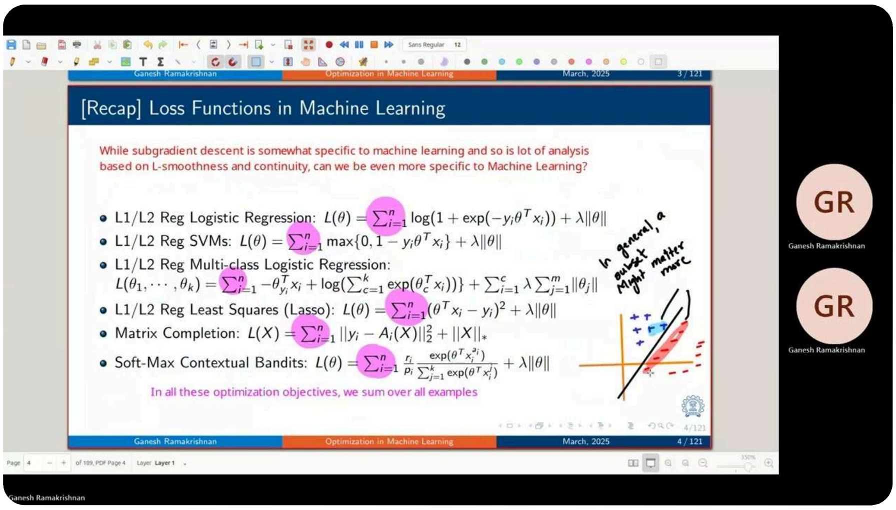
Advanced References and Optional Exercises:

  - Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
  - Optional Problem Solving: Understand the general implications of strong convexity on convergence on on page numbers 35-36 (sldes 179-180) of these notes
  - Best convergence rate bound for gradient (steepest) descent with constant step-size from these notes
  - Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.

PDF **Lecture 17: Minibatch, Hybrid and Accelerated Stochastic Gradient Descent**

Mark as done

Link to video:



Notes for iteration, editing, and discussion by students for Lecture 17 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/18-2FG-rTn5PchrdSKxKMFHTmUY0t2IM20cQewK0w8qA/edit?tab=t.0

Homework

- Homework1 with respect to pages 15-19 of the notes: Reflect and articulate the bias variance tradeoff between minibatch and vanilla SGD from the perspectives of (i) Reduced variance (ii) smaller difference between the Lipschitz continuity params B of SGD and GD and (iii) more meaningful batch compositions through diverse points in each set
  - Hint: Understand the landscape and practical implications of the comparison of Gradient Descent and Stochastic Gradient Descent on several fronts including convergence rates on page 16 of this deck
- Homework2 with respect to pages 22-25 of these notes: Despite the cons of stochastic gradient descent (SGD), why do ML researchers prefer SGD and its variants?

References (for optional reading):

- An overview of gradient descent optimization algorithms
- For subgradient methods, also refer to Section 1, 2 and 3 of these notes on advanced optimization by Boyd.
- Chapter on Stochastic Gradient Descent
- Blog on Stochastic Gradient Descent
- Identity Crisis: Memorization and Generalization under Extreme Overparameterization ICLR 2020: Link to slides on moodle and link to original paper. This came up in the discussion of solution to Homework from previous class.
- Reference: ADAM - 2015 ICLR paper : https://arxiv.org/pdf/1412.6980.pdf
- Reference: Stochastic Average Gradient 2016 paper: https://arxiv.org/pdf/1309.2388.pdf
- Reference: Unified Approach to Adaptive Regularization 2017 paper: https://arxiv.org/pdf/1706.06569.pdf and JMLR version at https://dl.acm.org/doi/10.5555/1953048.2021068 ;
- Link to a very nice blog: https://ruder.io/optimizing-gradient-descent/
- See https://ruder.io/optimizing-gradient-descent/index.html for a simple description of the evolution of, as well as of the strengths and weaknesses of different stochastic gradient (descent) algorithms.

Advanced References and Optional Exercises:

- Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.
- Online Gradient Descent: Efficient Algorithm for Regret Minimization - Zinkevich 2005
- Yu-Hong Dai, Roger Fletcher. New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds: http://link.springer.com/content/pdf/10.1007%2Fs10107-005-0595-2.pdf
- Regret bound for Adagrad: Duchi, Hazan, Singer 2010 as well as Kalai-Vempala 2005
- Some more advanced reading is here

Mark as done

Preparation for Lecture 18

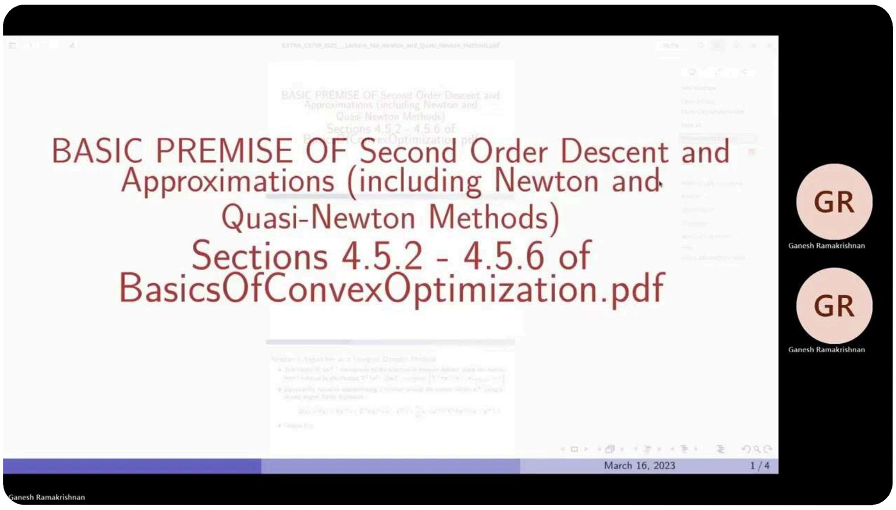
Mark as done

Preparation for Lecture 19

Lecture 18: ADAM, Adagrad, RMSProp and toward Proximal and Generalized Gradient Descent

Mark as done

Video Lecture:



Notes for iteration, editing, and discussion by students for Lecture 18 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1uY-Qt0i8ZsckiBae72x1_tUZlxrZbLzy2M40ov-7nTs/edit?tab=t.0

Homework

- Understand and internalize the unified framework for adaptive+accelerated variants of stochastic gradient descent on pages 8-11 of the slides.
- With the background provided on pages 8-11 of the slides appreciate in details, 1) the steps of the ADAM algorithm (which uses Heavy Ball Momentum) and 2) the steps of a similar NADAM algorithm (which uses Nestorov Accelerated Momentum) also presented in the class

References (for optional reading):

- Identity Crisis: Memorization and Generalization under Extreme Overparameterization ICLR 2020: Link to slides on moodle and link to original paper. This came up in the discussion of solution to Homework from previous class.
- Reference: ADAM - 2015 ICLR paper : https://arxiv.org/pdf/1412.6980.pdf
- Reference: Stochastic Average Gradient 2016 paper: https://arxiv.org/pdf/1309.2388.pdf
- Reference: Unified Approach to Adaptive Regularization 2017 paper: https://arxiv.org/pdf/1706.06569.pdf and JMLR version at https://dl.acm.org/doi/10.5555/1953048.2021068 ;
- Link to a very nice blog: https://ruder.io/optimizing-gradient-descent/
- See https://ruder.io/optimizing-gradient-descent/index.html for a simple description of the evolution of, as well as of the strengths and weaknesses of different stochastic gradient (descent) algorithms.

Advanced References and Optional Exercises:

- Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov
- Online Gradient Descent: Efficient Algorithm for Regret Minimization - Zinkevich 2005
- Yu-Hong Dai, Roger Fletcher. New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds: http://link.springer.com/content/pdf/10.1007%2Fs10107-005-0595-2.pdf
- Regret bound for Adagrad: Duchi, Hazan, Singer 2010 as well as Kalai-Vempala 2005

- Some more advanced reading is here

17 March - 23 March

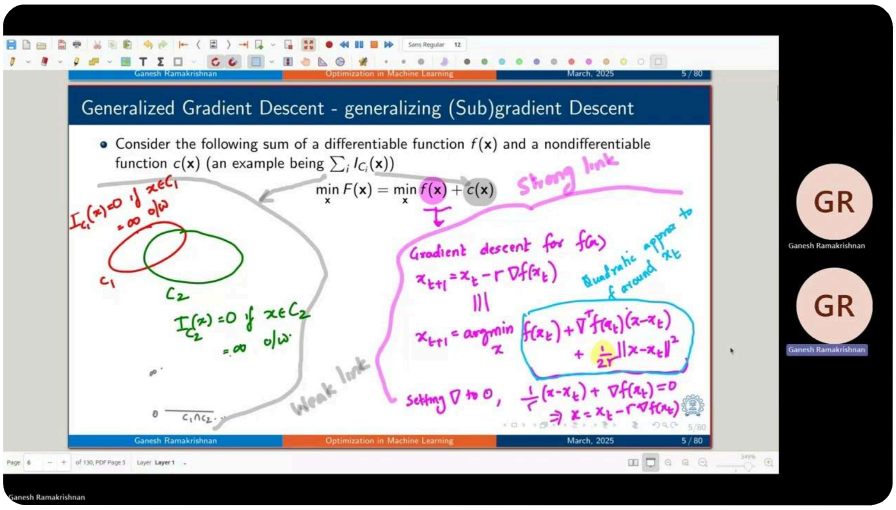
24 March - 30 March

31 March - 6 April

Lecture 19: Proximal, Generalized Gradient Descent and its Acceleration and Projection in GD

Mark as done

Link to video lecture:



Notes for iteration, editing, and discussion by students for Lecture 19 [Students encouraged to add thoughts/including equations or doubts/questions]: https://docs.google.com/document/d/1_TUDkfjgcfCQbIkjMiZmo8VFO_EbONPDVx8jbzSSfsg/edit?tab=t.0

Homework

- Recall the Sufficient condition test for simplified lasso. Verify that plugging it in the proximal operator therein yields the detailed steps of the overall Iterative Soft Thresholding Algorithm (ISTA) for LASSO. Also, contemplate the convergence of the ISTA algorithm (see pages 17-22 of the slide deck)

References (for optional reading):

- See https://ruder.io/optimizing-gradient-descent/index.html for a simple description of the evolution of, as well as of the strengths and weaknesses of different stochastic gradient (descent) algorithms.
- For Projected and Generalized Gradient Descent, see L. Vandenberghe, Lecture Notes for EE 236C.
- Reference: Page 557, Problem 10.2 of Boyd and Vandenberghe
- Reference: Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.
- References: For solutions to SVM using different algorithms referred to in the class, including projected gradient descent, refer here.

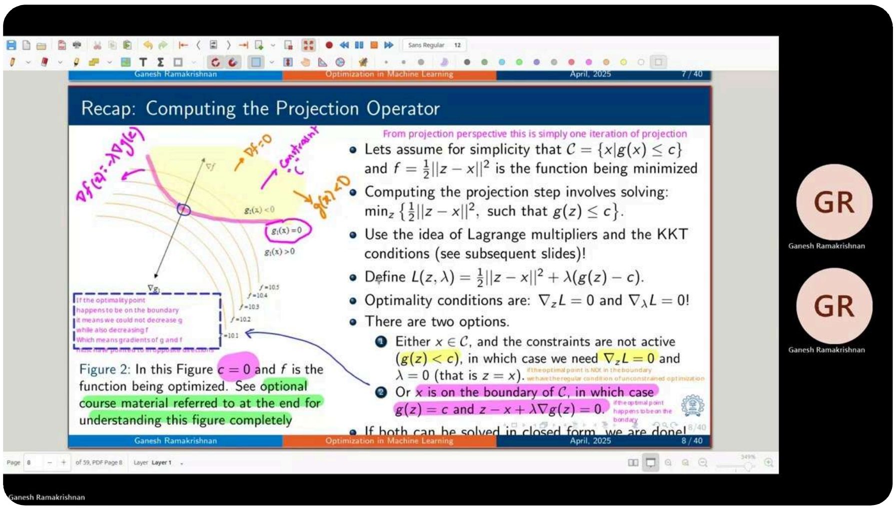
Advanced References and Optional Exercises:

- Online Gradient Descent: Efficient Algorithm for Regret Minimization - Zinkevich 2005
- Yu-Hong Dai, Roger Fletcher. New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds: http://link.springer.com/content/pdf/10.1007%2Fs10107-005-0595-2.pdf
- Regret bound for Adagrad: Duchi, Hazan, Singer 2010 as well as Kalai-Vempala 2005
- Some more advanced reading is here

7 April - 13 April

[Lecture 20: Project Gradient Descent, Karush Kuhn Tucker Conditions and Duality, Further results in Discrete Optimization]() Mark as done

Link to video lecture:



Notes for iteration, editing, and discussion by students for Lecture 20 [Students encouraged to add thoughts/including equations or doubts/questions]:https://docs.google.com/document/d/1rbnWSZU497o1aHdU3bpzohn8h-KlbEdQgXZKCSDDymo/edit?tab=t.0

Homework

- HOMEWORK: CAN THE FORM OF THE DUAL BECOME SIGNIFICANTLY SIMPLER THAN THE PRIMAL IS THAT WHAT HAPPENED IN SVMs on page 47? Is the Dual always guaranteed to be a concave maximization problem even if the primal is not a convex problem to minimize.
- Homework: On page 90, try and complete the proof that Union Intersection property of Submodular functions implies the diminishing gains property.

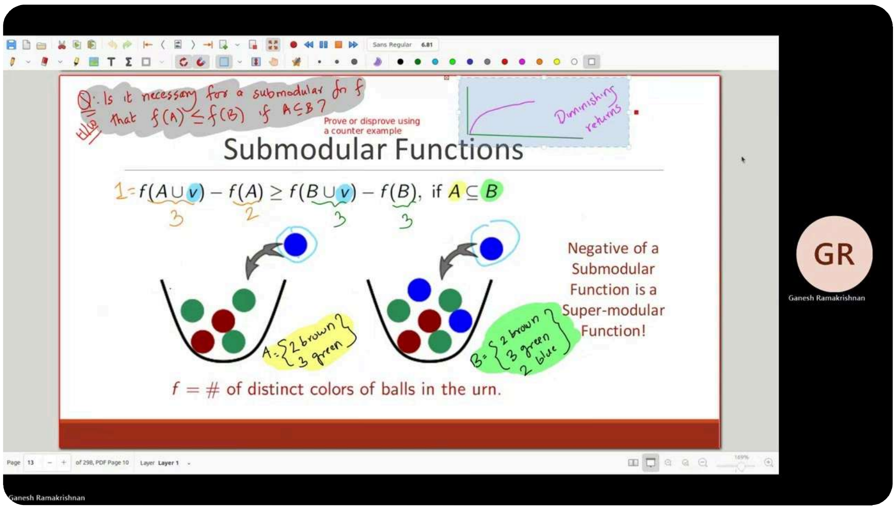
References (for optional reading):

- References: Boyd and Vandenberghe Sections 5.1, 5.2. Geometric motivation for lagrange multipliers and KKT necessary conditions in Section 4.4.1 and 4.4.2 (page 282 onwards) of [my lecture notes]()
- References: For Proximal and Project Operator, see [Chapter 6 of book First-Order Methods in Optimization by Amir Beck]().
- References: For Proximal Gradient Algorithm, see [Chapter 10 of book First-Order Methods in Optimization by Amir Beck]().
- References: For Projected and Generalized Gradient Descent, see [L. Vandenberghe, Lecture Notes for EE 236C](). Some advanced reading is [here]()
- Reference: Page 557, Problem 10.2 of Boyd and Vandenberghe

Advanced References and Optional Exercises:

- Reference: Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of [Introductory Lectures on Convex Programming]() by Nesterov.
- Fujishige, “Submodular Functions and Optimization”, 2005
- Narayanan, “Submodular Functions and Electrical Networks”, 1997
- Welsh, “Matroid Theory”, 1975
- Oxley, “Matroid Theory”, 1992 (and 2011).
- Lawler, “Combinatorial Optimization: Networks and Matroids”, 1976.
- Schrijver, “Combinatorial Optimization”, 2003
- Gruenbaum, “Convex Polytopes, 2nd Ed”, 2003.

[[Optional material for self-learning] Lecture 22: Submodular functions, equivalent definitions, basic properties]() Mark as done



Homework

- Can you try and prove that the two definitions of submodular functions on page 30 are equivalent?
- On pages 31 and 32, we discuss the max (and min) of submodular functions. Using any hints available therein, what can you claim about the max and min of two submodular functions? Are either or both or neither of them submodular. How can you prove your answer?

References (for optional reading):

- Reference: Andreas Krause's web page: http://submodularity.org
- Reference: Slides+Videos from our tutorial at https://sites.google.com/view/ijcaitutorial2020summarization/home ;

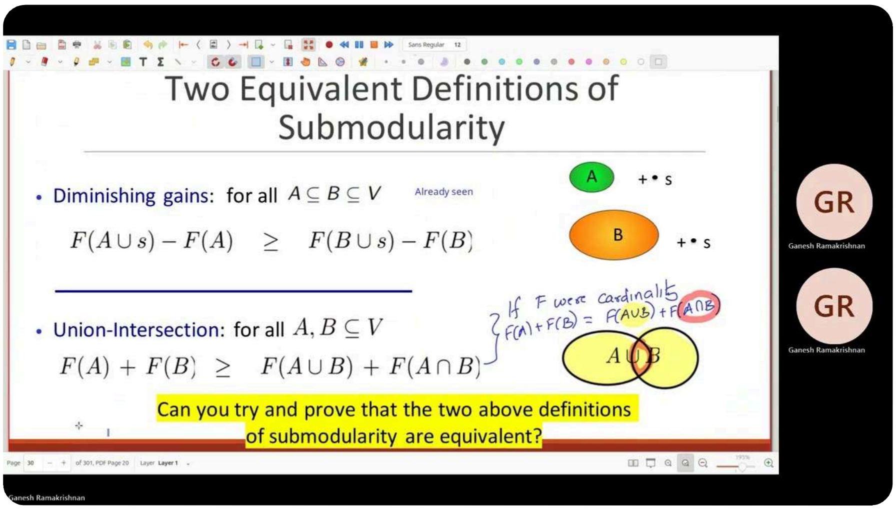
Advanced References and Optional Exercises:

- Reference: Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.
- Fujishige, "Submodular Functions and Optimization", 2005
- Narayanan, "Submodular Functions and Electrical Networks", 1997
- Welsh, "Matroid Theory", 1975
- Oxley, "Matroid Theory", 1992 (and 2011).
- Lawler, "Combinatorial Optimization: Networks and Matroids", 1976.
- Schrijver, "Combinatorial Optimization", 2003
- Gruenbaum, "Convex Polytopes, 2nd Ed", 2003.

Lecture 21: Submodular functions, equivalent definitions, basic properties

Mark as done

Link to video lecture:



Homework

- On pages 34-37, we discuss the max (and min) of submodular functions. Using any hints available therein, what can you claim about the max and min of two submodular functions? Are either or both or neither of them submodular. How can you prove your answer?

References (for optional reading):

- Reference: Andreas Krause's web page: http://submodularity.org

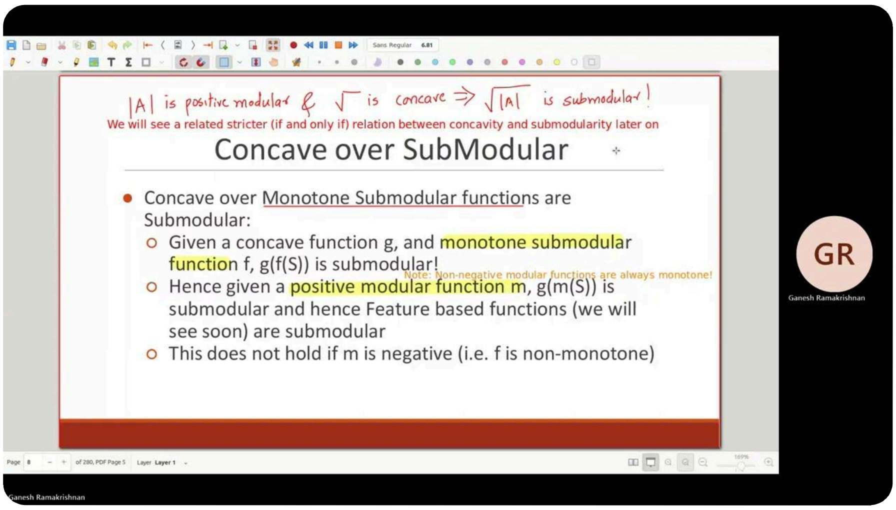
- Reference: Slides+Videos from our tutorial at https://sites.google.com/view/ijcaitutorial2020summarization/home ;

Advanced References and Optional Exercises:

- Reference: Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.
- Fujishige, “Submodular Functions and Optimization”, 2005
- Narayanan, “Submodular Functions and Electrical Networks”, 1997
- Welsh, “Matroid Theory”, 1975
- Oxley, “Matroid Theory”, 1992 (and 2011).
- Lawler, “Combinatorial Optimization: Networks and Matroids”, 1976.
- Schrijver, “Combinatorial Optimization”, 2003
- Gruenbaum, “Convex Polytopes, 2nd Ed”, 2003.

[Extra and Optional: Previous Year] Lecture 22: Examples of Submodular functions: Cooperative costs, Attractive Potentials, Complexity, Representation, Diversity, Coverage

Mark as done



References (for optional reading):

- Reference: Andreas Krause’s web page: http://submodularity.org
- Reference: Slides+Videos from our tutorial at https://sites.google.com/view/ijcaitutorial2020summarization/home ;

Advanced References and Optional Exercises:

- Reference: Sections 3.2, 3.2.3, 3.2.4 (Advanced) and 3.2.5 (Advanced) of Introductory Lectures on Convex Programming by Nesterov.
- Fujishige, “Submodular Functions and Optimization”, 2005
- Narayanan, “Submodular Functions and Electrical Networks”, 1997
- Welsh, “Matroid Theory”, 1975
- Oxley, “Matroid Theory”, 1992 (and 2011).
- Lawler, “Combinatorial Optimization: Networks and Matroids”, 1976.
- Schrijver, “Combinatorial Optimization”, 2003
- Gruenbaum, “Convex Polytopes, 2nd Ed”, 2003.
