Optimization

Hongyuan Mei
JHU-CLSP

Model

- What is Model
  - A set of assumptions over data
  - The distribution constructed over data
    - Defined based on the set of assumptions

Model

- Observe a sequence of HHTTTH…
- To learn $P(X = H) = \theta_1$ and $P(X = T) = \theta_2$
  - $X_n \sim P_\theta(X)$ where they are i.i.d.
  - Probability of any sequence of length N
  - $P_\theta(X_1, X_2, \ldots, X_N) = \prod_{n=1}^{N} P_\theta(X_n)$

Likelihood

- $P_\theta(X_1, X_2, \ldots, X_N) = \prod_{n=1}^{N} P_\theta(X_n)$
- *likelihood function* of θ given observations $\{X_n\}_{n=1}^{N}$
- We want our model to well explain the data
- We *estimate* θ by *maximizing likelihood*
- So $\hat{\theta}$ is *Maximum Likelihood Estimate*

Cross-Entropy

- We work in log-space to prevent underflow
- $\log P_\theta(X_1, X_2, \ldots, X_N) = \sum_{n=1}^{N} \log P_\theta(X_n)$
- *log-likelihood function*
- Maximize log-likelihood — Minimize cross-entropy
- Or *loss* $\ell = -\sum_{n=1}^{N} \log P_\theta(X_n)$

Optimization

- Find $\hat{\theta}$ to minimize loss $\ell = -\sum_{n=1}^{N} \log P_\theta(X_n)$
- We call it *optimization*
- But HOW?
- Is there any properties that makes our life easier?

Convexity

- A set C is *convex* if and only if
  - $\forall a, b \in C$, $\lambda a + (1 - \lambda) b \in C$ for all $\lambda \in (0, 1)$
- Line, segment, circles, half-plane, etc

Convexity

- A function $f(x)$ is *convex* on C if and only if
  - $\forall x_1, x_2 \in C$ and $\lambda \in (0, 1)$
  - $f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$
- Strictly convex if
  - $f(\lambda x_1 + (1 - \lambda) x_2) < \lambda f(x_1) + (1 - \lambda) f(x_2)$

Convex Optimization

- EASY to minimize convex function on convex set
  - Any local minimum is a global minimum.
  - Strictly convex — at most one global minimum.
  - Solve by setting first-order derivative to 0
    - $df(x)/dx = 0$ for scalar x
    - $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0}$ for vector $\mathbf{x}$

Convex Functions

- Commonly used convex functions
  - Affine $f(x) = ax + b$
  - Quadratic $f(x) = ax^2 + bx + c$ where $a > 0$
  - Negative logarithm $f(x) = -log(x)$
  - Sum of convex functions!

Unconstrained

- One example:
  - $f(x) = x^2 - 4x + 5$
  - $df(x)/dx = 2x - 4 = 0$
  - $x = 2$

Coin Flip

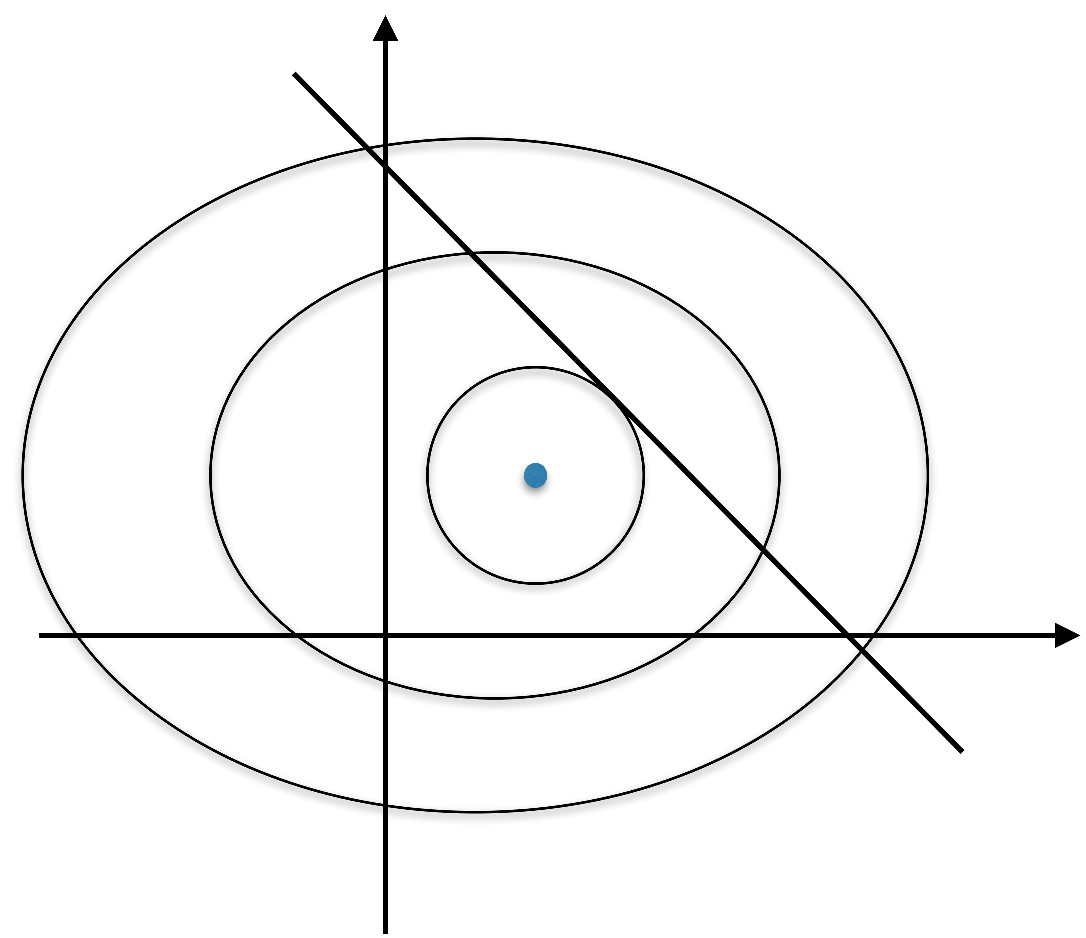
- Observe a sequence of HHTTTH…
- $N = 100 \qquad c(H) = 46 \qquad c(T) = N - c(H) = 54$
- So…
- $\ell = -\sum_{n=1}^{N} \log P_\theta(X_n)$
- $\ell = -c(H)\log P(H) - c(T)\log P(T)$
- $\ell = -c(H)\log \theta_1 - c(T)\log \theta_2$

Coin Flip

- $\partial\ell/\partial\theta_1 = -c(H)/\theta_1$ and $\partial\ell/\partial\theta_2 = -c(T)/\theta_2$
- $\theta_1 = \theta_2 = \infty$?
- Probabilities can NOT go wildly!
- $\theta_1 + \theta_2 = 1$
- So it is a constrained optimization problem
- How do we deal with it?

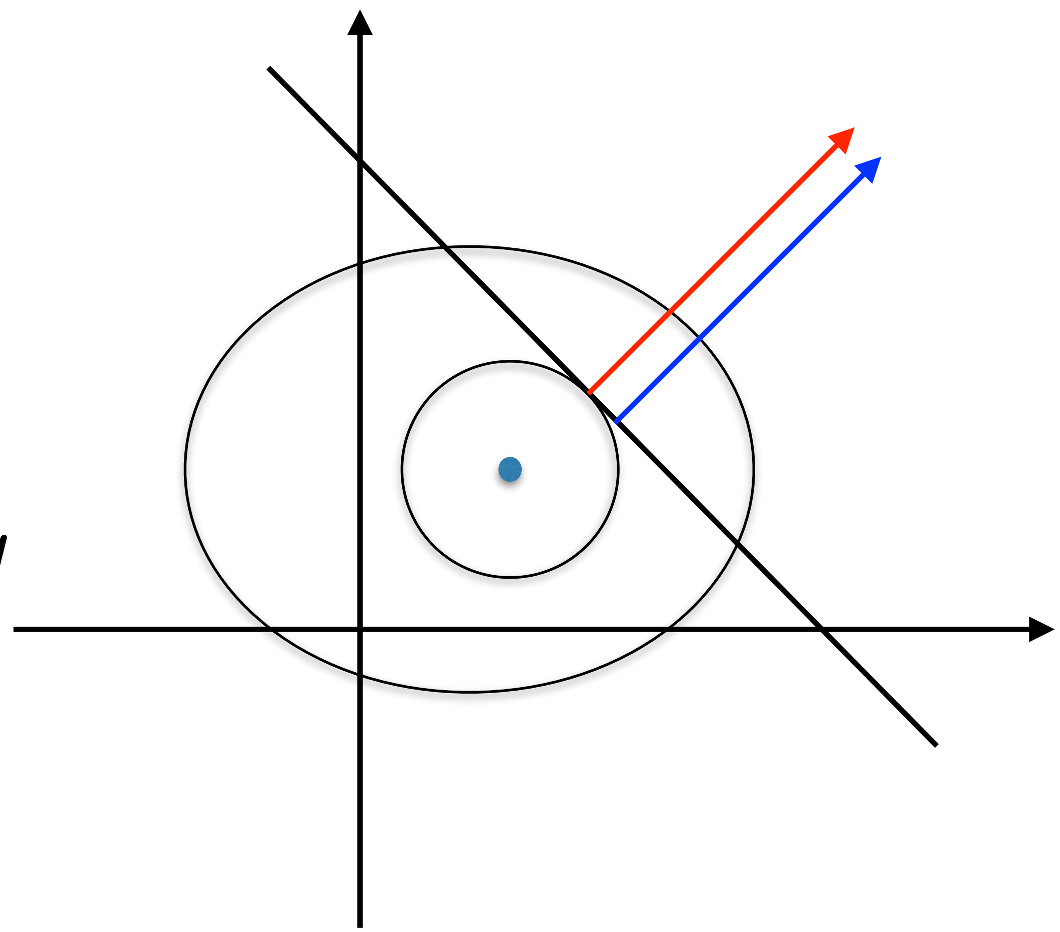
Constrained

- $\ell = -c(H)\log\theta_1 - c(T)\log\theta_2$
- $g = \theta_1 + \theta_2 - 1$
- Optima at tangent point
  - $\nabla\ell = \lambda\nabla g$
  - $g = 0$



Lagrangian

- Optima at tangent point
  - Why $\nabla \ell = \lambda \nabla g$?
  - Normal vector of ℓ — $\nabla \ell$
  - Normal vector of g — ∇g
  - Parallel to each other



Lagrangian

- Constrained —> Unconstrained
- $\mathcal{L} = \ell - \lambda g$ with parameters $\theta_1, \theta_2, \lambda$
  - $\nabla \mathcal{L} = \nabla \ell - \lambda \nabla g = 0$ i.e. $\nabla \ell = \lambda \nabla g$
  - $\partial \mathcal{L} / \partial \lambda = -g = 0$ i.e. $g = 0$
- Converted by introducing *Lagrangian Multiplier*
- Still convex!

Coin Flip

- $\mathcal{L} = -c(H)\log\theta_1 - c(T)\log\theta_2 - \lambda(\theta_1 + \theta_2 - 1)$
- $\partial\mathcal{L}/\partial\theta_1 = -c(H)/\theta_1 - \lambda = 0$
- $\partial\mathcal{L}/\partial\theta_2 = -c(T)/\theta_2 - \lambda = 0$
- $\partial\mathcal{L}/\partial\lambda = -\theta_1 - \theta_2 + 1 = 0$
- Now, let's do high school review
- $\lambda = -N \quad \theta_1 = c(H)/N \quad \theta_2 = c(T)/N = 1 - c(H)/N$

Recitation Question

- Recitation Loglin 1(a)
- Bwa = 1, Bwee = 2, Kiki = 3
- Observe $c(1)/N = 0.3 \;\; c(2)/N = 0.20 \;\; c(3)/N = 0.5$
- $\ell = -c(1)\log\theta_1 - c(2)\log\theta_2 - c(3)\log\theta_3$
- $\theta_1 + \theta_2 + \theta_3 = 1$
- Now do your exercise!

Recitation Question

- $\mathcal{L} = \ell - \lambda g$
- $\ell = -c(1)\log\theta_1 - c(2)\log\theta_2 - c(3)\log\theta_3$
- $g = \theta_1 + \theta_2 + \theta_3 - 1$
- $\forall i \in \{1,2,3\} \quad \partial\mathcal{L}/\partial\theta_i = -c(i)/\theta_i - \lambda = 0$
- $\partial\mathcal{L}/\partial\lambda = -\theta_1 - \theta_2 - \theta_3 + 1 = 0$
- $\lambda = -c(1) - c(2) - c(3) = -N$ and $\theta_i = c(i)/N$

Inequality

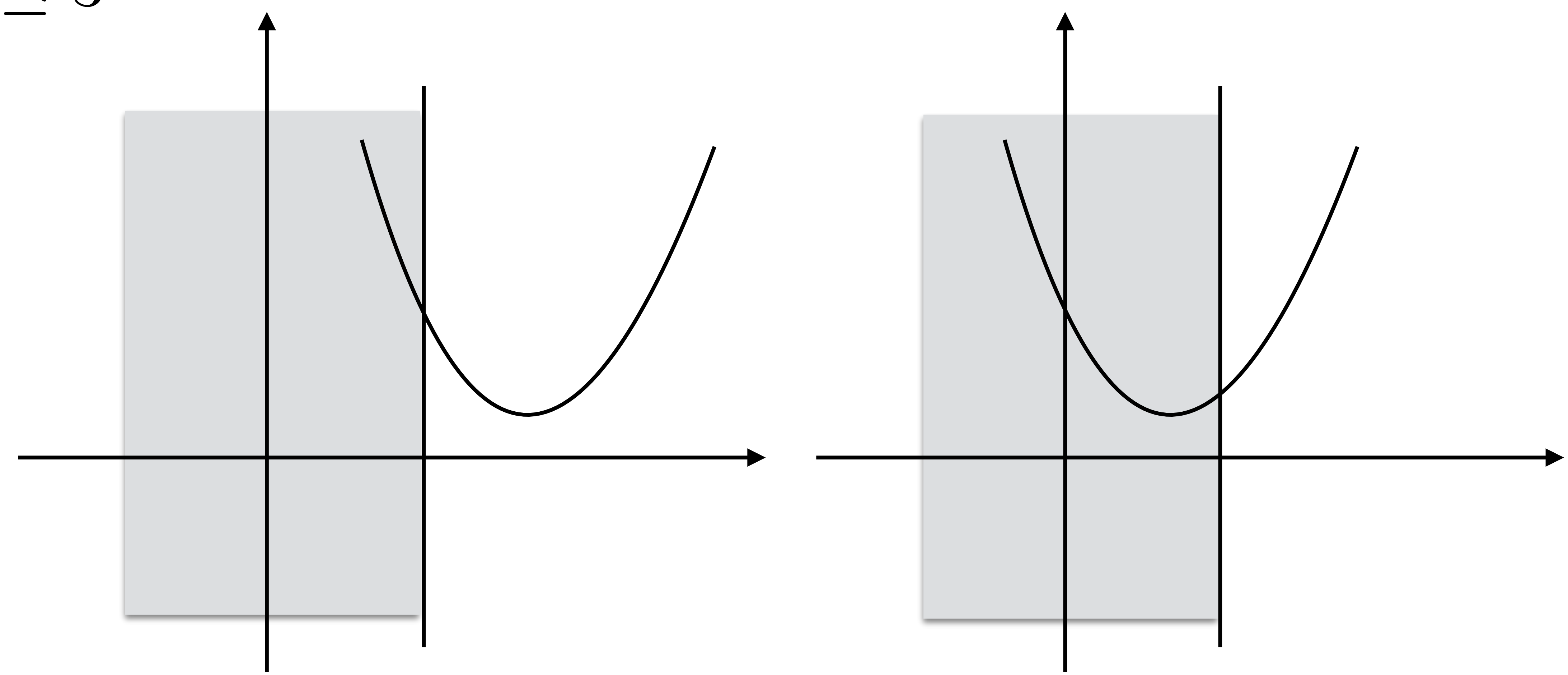
- Wait! We may forget something…
- $\forall i \in \{1,2,3\} \quad 0 \leq \theta_i \leq 1$
- In NLP class, no worry about this
- But let's go beyond!

Inequality

- One example:
  - $f(x) = x^2 - 4x + 5$
  - $x \leq 3$
  - Pretend to NOT know the answer…

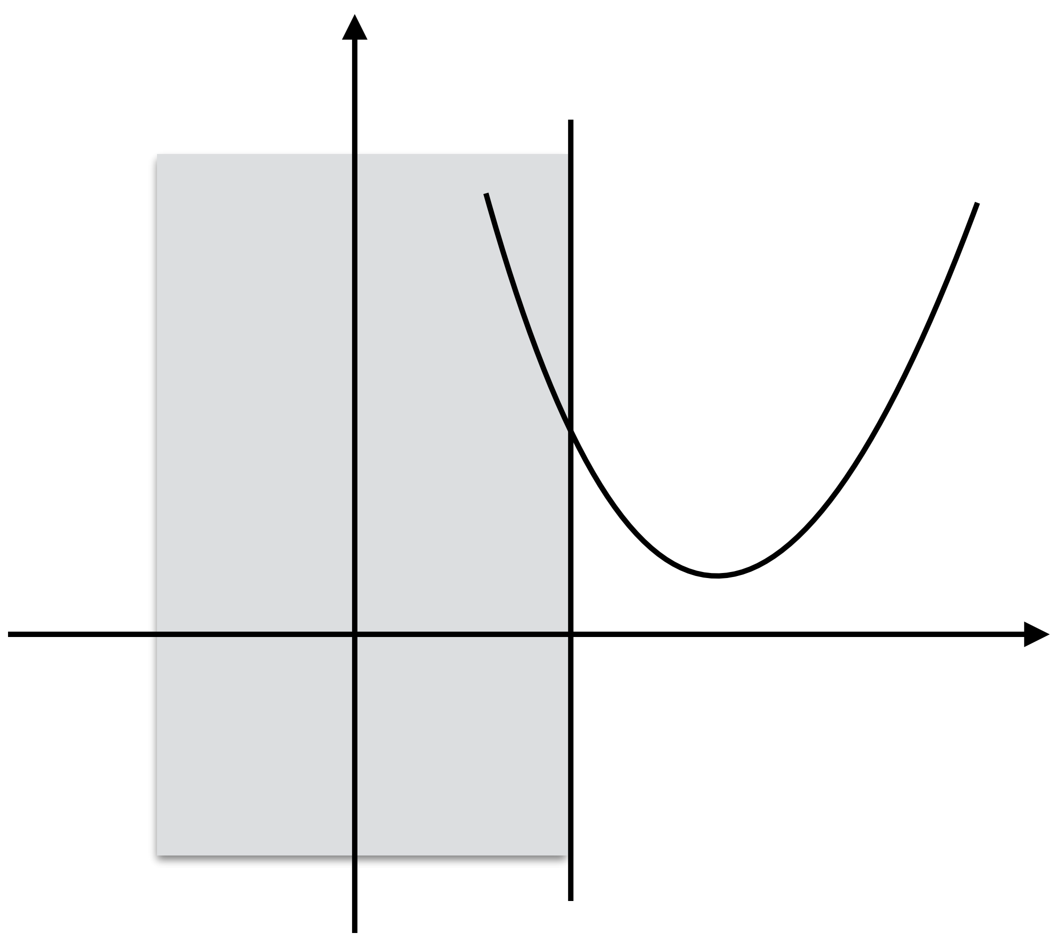
Inequality

- $f(x) = x^2 - 4x + 5$
- $x \leq 3$

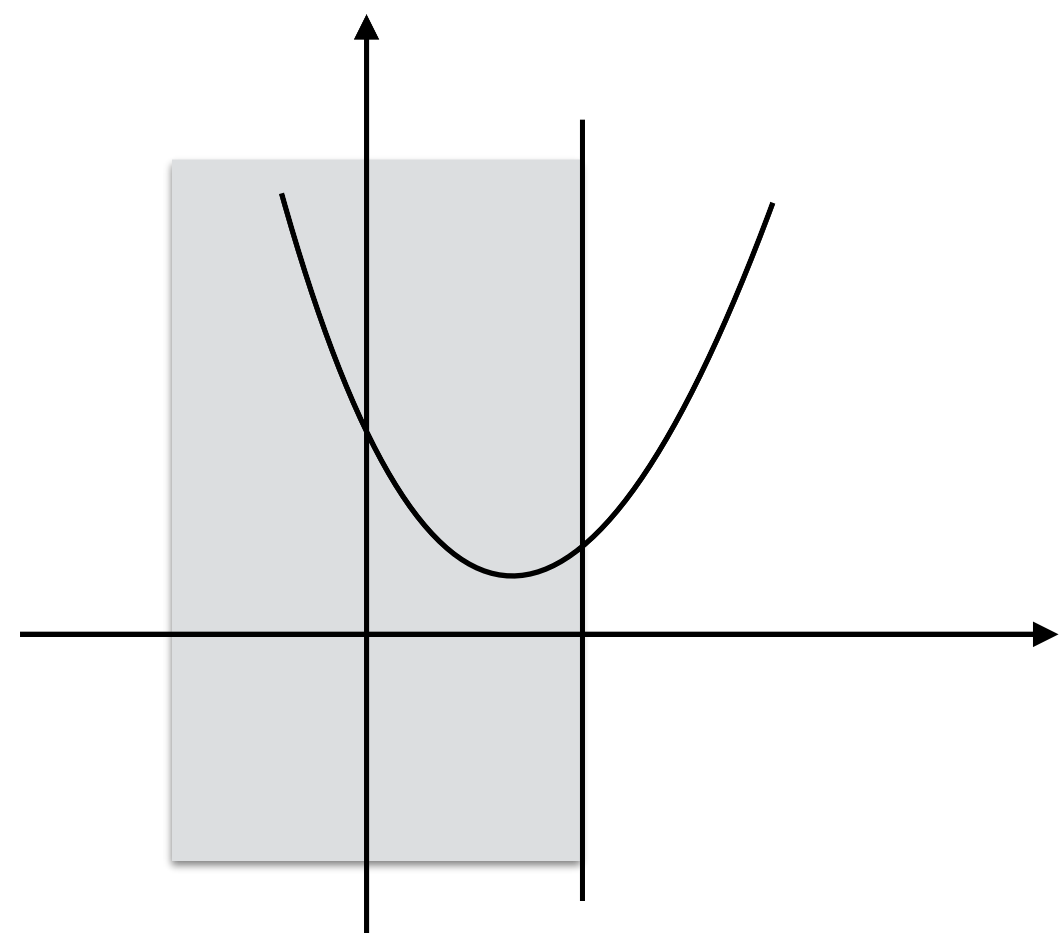


Inequality

- $x^* > 3$
- At the boundary!

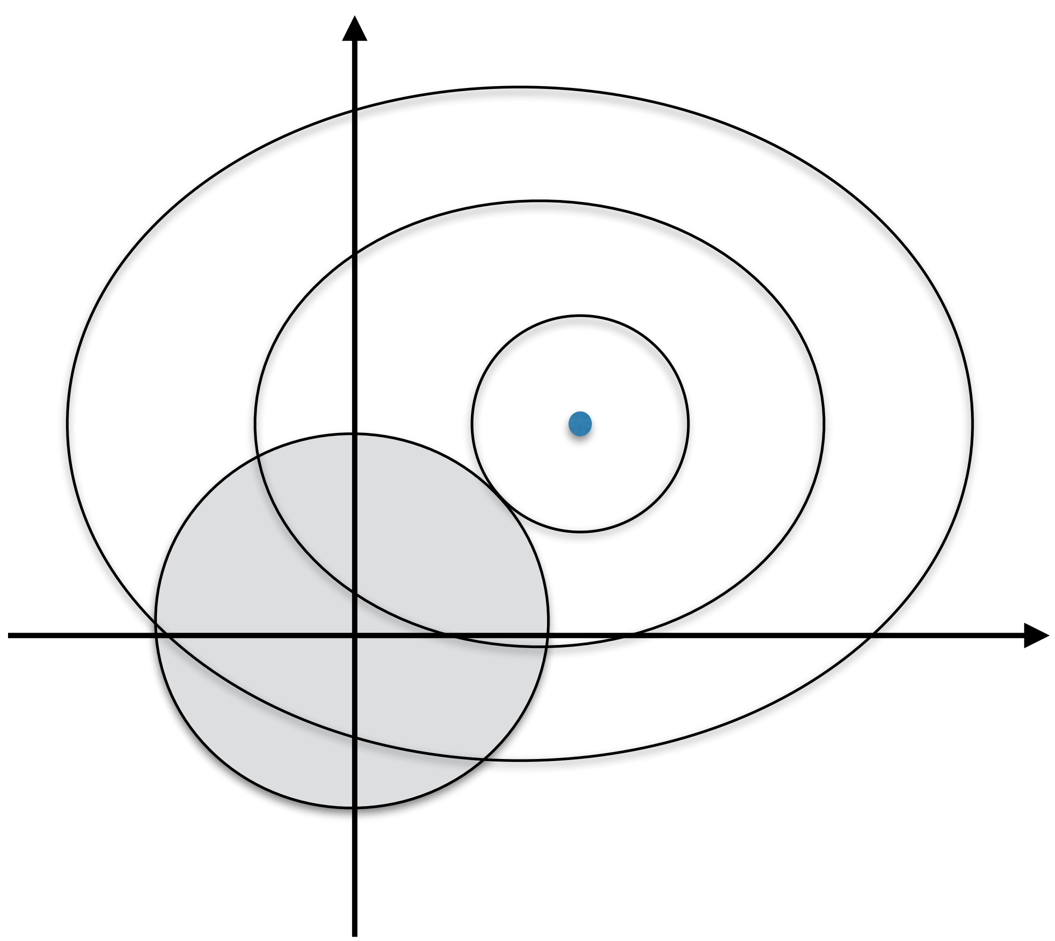


- $x^* \leq 3$
- Unconstrained again!

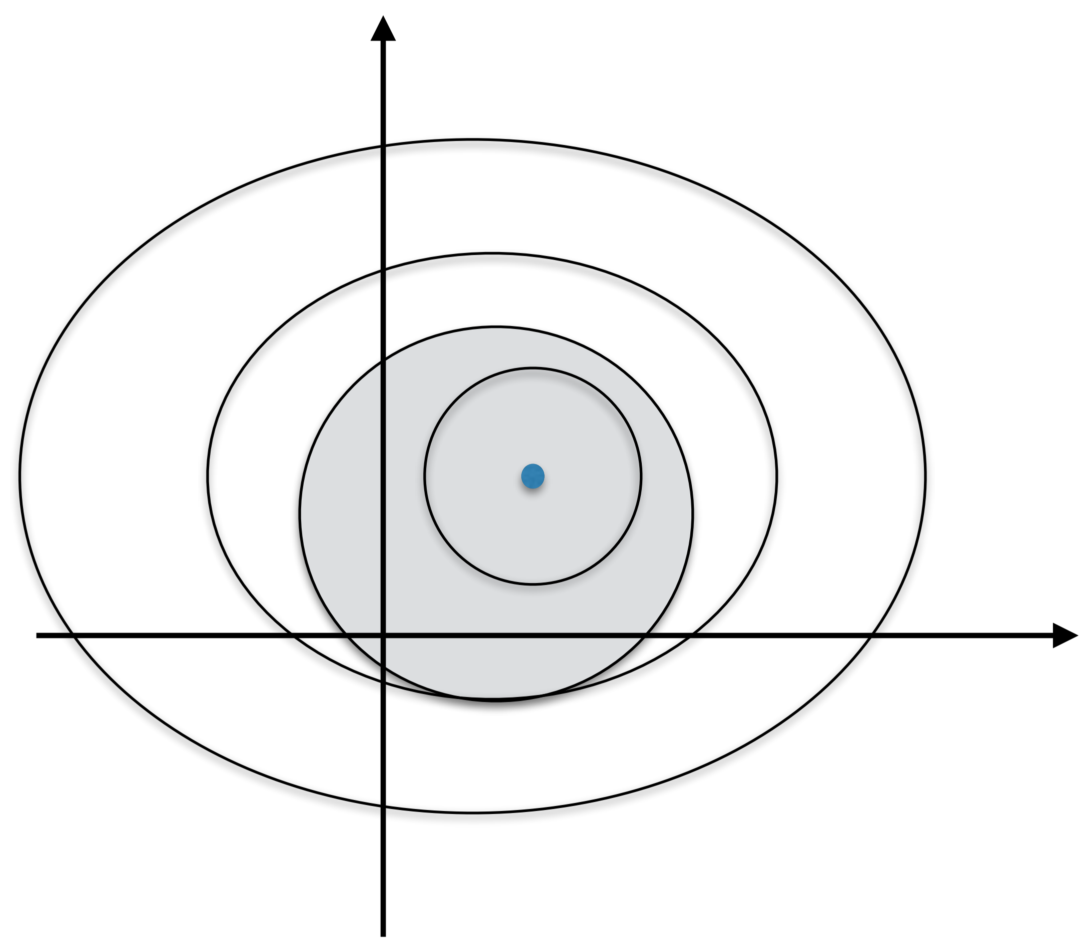


Another Example

- Optima out of constraints
- At the boundary!

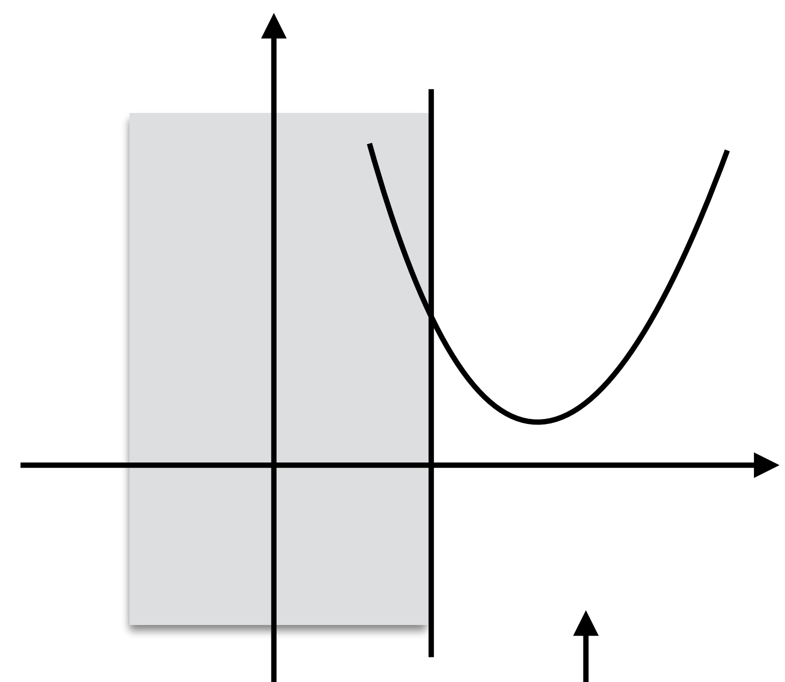


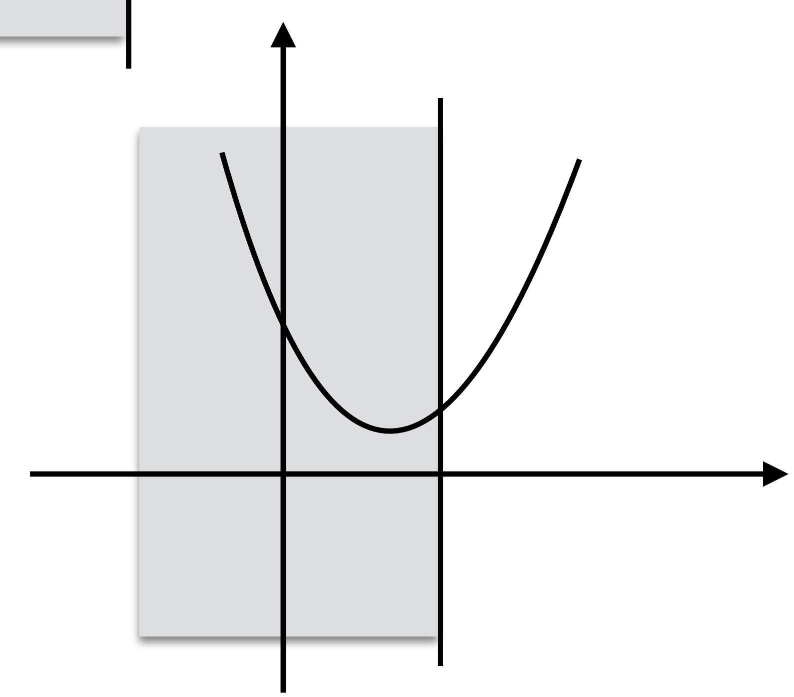
- Optima under constraints
- Unconstrained again!



Conversion

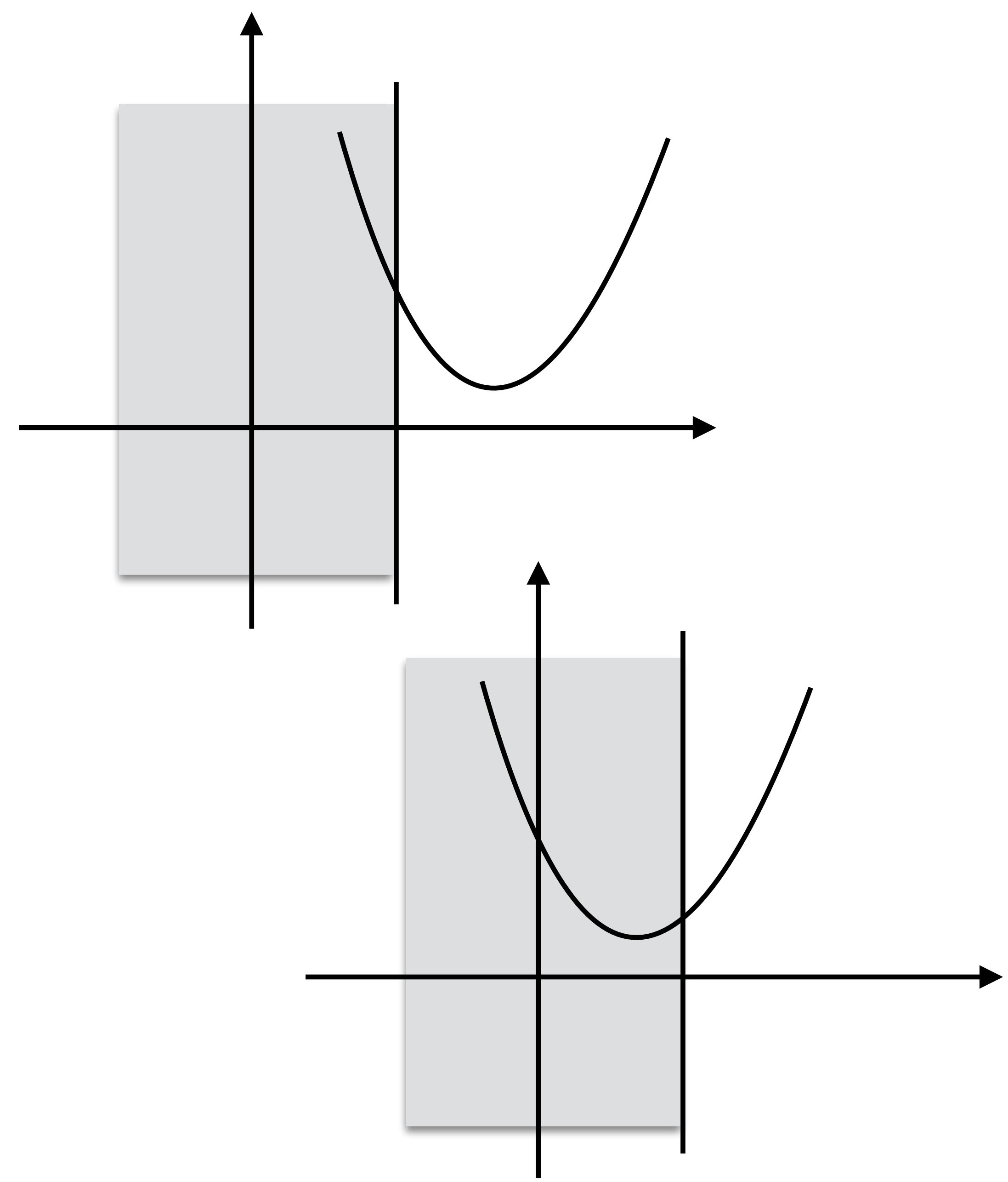
- $\mathcal{L} = f(x) + \mu g(x)$
- $f(x) = x^2 - 4x + 5$
- $g(x) = x - 3$
- What we can say for x^*, μ^*
  - $\nabla f(x^*) + \mu^* \nabla g(x^*) = 0$
  - $g(x^*) \leq 0$
  - And what?





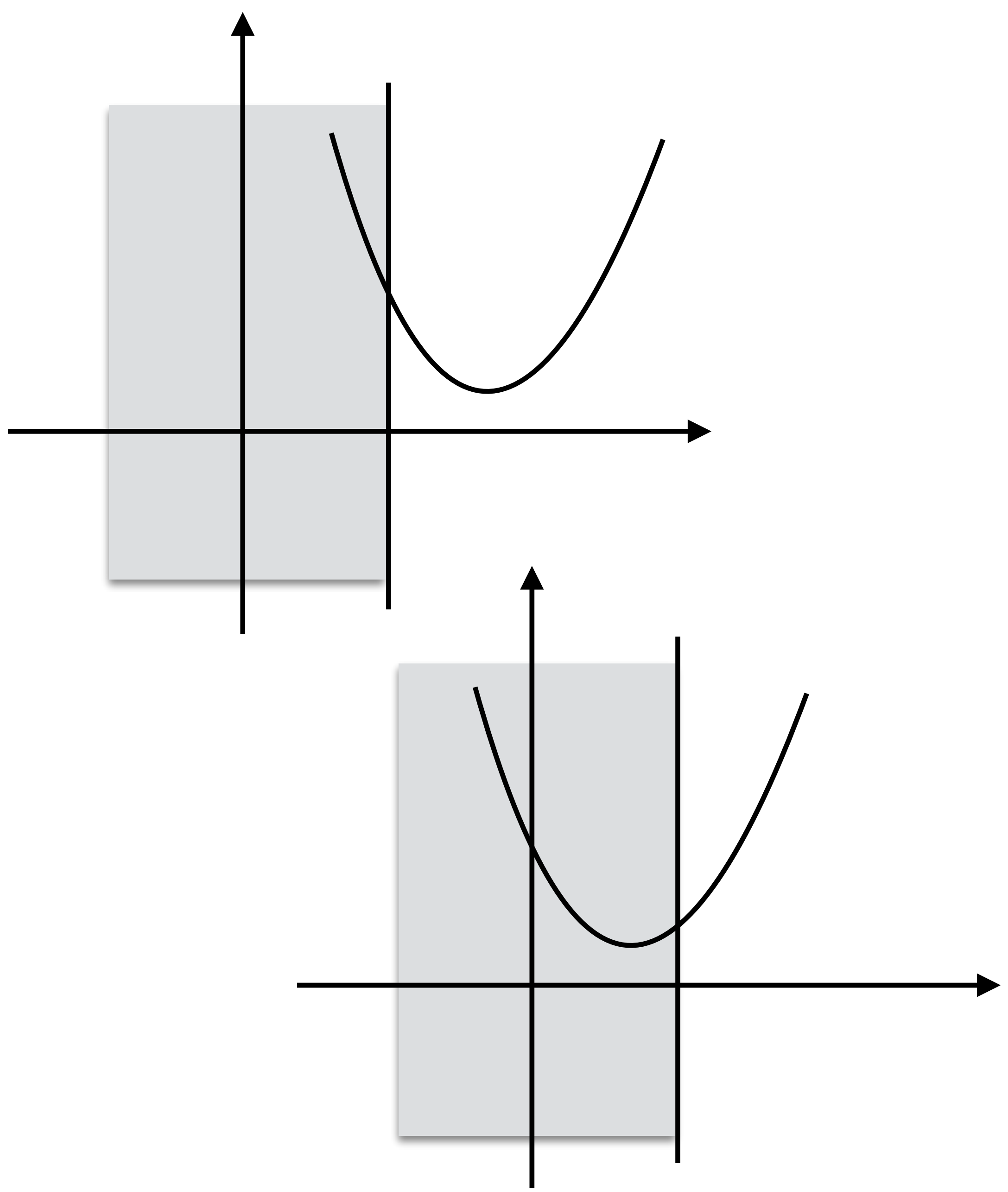
Conversion

- $\mu^* g(x^*) = 0$ and $\mu^* \geq 0$
- Why?
- If $\mu^* > 0$
    - $g(x^*) = 0$ (boundary!)
    - Otherwise $\min \mathcal{L} = -\infty$
- If $\mu^* = 0$
    - Unconstrained again!



KKT Conditions

- $\nabla f(x^*) + \mu^* \nabla g(x^*) = 0$
- $g(x^*) \leq 0$
- $\mu^* g(x^*) = 0$
- $\mu^* \geq 0$
- KKT conditions



KKT Conditions

- $\nabla f(x^*) + \mu^* \nabla g(x^*) = 0$
- $g(x^*) \leq 0$
- $\mu^* g(x^*) = 0$
- $\mu^* \geq 0$
- KKT conditions

- $2x^* - 4 + \mu^* = 0$ (1)
- $x^* \leq 3$ (2)
- $\mu^*(x^* - 3) = 0$ (3)
- $\mu^* \geq 0$ (4)
- (1)(3) —> $\mu^* \in \{0, -2\}$
- (4) —> $\mu^* = 0$
- (1)(2) —> $x^* = 2$

KKT Conditions

- $\nabla f(x^*) + \mu^* \nabla g(x^*) = 0$
- $g(x^*) \leq 0$
- $\mu^* g(x^*) = 0$
- $\mu^* \geq 0$
- KKT conditions

- They are very important
- In machine learning
- Meet KKT conditions
- Close *duality gap*
- Out of our scope for now
- But crucial for SVM

Non-Convex

- What if NON-convex?
- SGD or EM
- Next lecture (when Jason start talking about EM)
  - He will show you guys a really nice example!

Thanks!

Special Acknowledgement to Xiaochen Li

Appendix

- Sum of convex functions
    - My notes
    - https://docs.google.com/document/d/1yA2obUuxXcxC84U1D4lfSyijFthHDUnkKNy3_FrwtCQ/edit?usp=sharing