CSCI B609: "Foundations of Data Science"

Lecture 10/11: Random Walks and Markov Chains

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

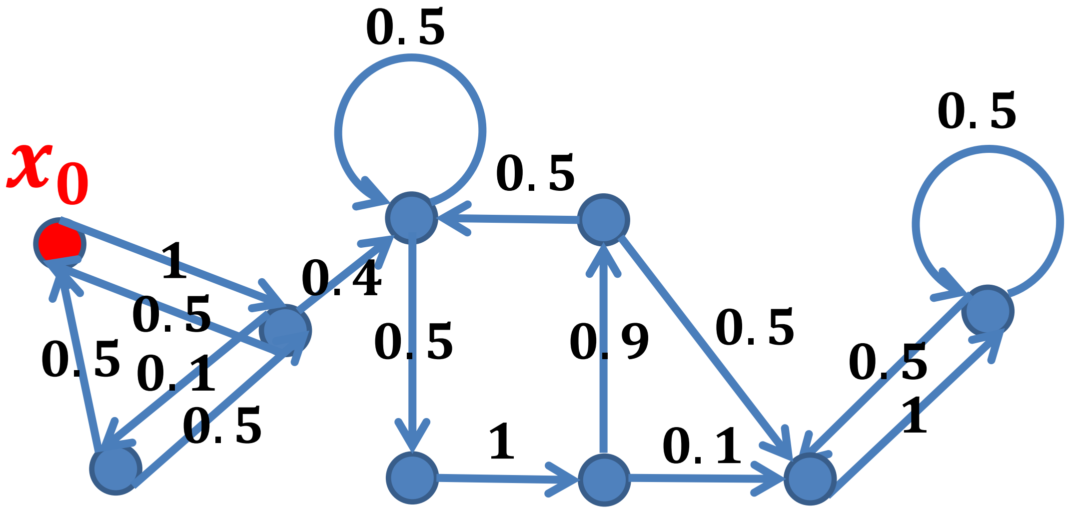
http://grigory.us

Project Example: Gradient Descent in TensorFlow

- Gradient Descent (will be covered in class)
- Adagrad: http://www.magicbroom.info/Papers/DuchiHaSi10.pdf
- Momentum (stochastic gradient descent + tweaks): http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf
- Adam (Adaptive + momentum): http://arxiv.org/pdf/1412.6980.pdf
- FTRL: http://jmlr.org/proceedings/papers/v15/mcmahan11b/mcmahan11b.pdf
- RMSProp: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

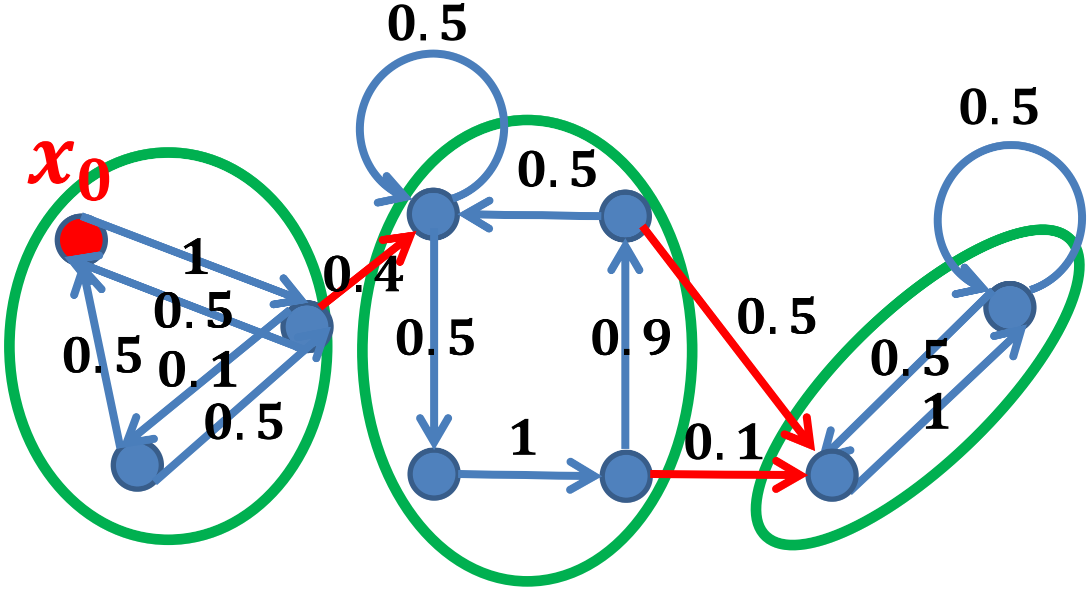
Random Walks and Markov Chains

- Random walk:
  - Directed graph $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$
  - Starting vertex $\boldsymbol{x_0} \in \boldsymbol{V}$
  - Edge (i,j): probability p_{ij} of transition $i \to j$
  - $\forall i: \sum_j p_{ij} = 1$



Strongly Connected Components

- **Def (Strongly Connected Component).** $S \subseteq V$ such that $\forall i, j \in S$ there exist paths $i \to j$ and $j \to i$
- SCC's form a partition of the vertex set
- **Terminal SCC**: no outgoing edges
- Long enough random walk $\to$ **Terminal SCC**



Matrix Form and Stationary Distribution

- $\boldsymbol{p}_{\boldsymbol{t}} =$ probability distribution over vertices at time $\boldsymbol{t}$
- $\boldsymbol{p}_0 = (1,0,0, \dots, 0)$
- $\boldsymbol{p}_{\boldsymbol{t}} P = \boldsymbol{p}_{\boldsymbol{t+1}}$
- P = transition matrix with entries p_{ij}
- If $\boldsymbol{t} \to \infty$ then average of $\boldsymbol{p}_i's$ converges:

$$\frac{1}{\boldsymbol{t}} \sum_{\boldsymbol{i}=0}^{\boldsymbol{t-1}} \boldsymbol{p}_{\boldsymbol{i}} \to \boldsymbol{\pi}$$

- $\boldsymbol{\pi} =$ **stationary distribution** of P
- $\boldsymbol{\pi}$ is unique and doesn't depend on x_0 if G is strongly connected
- Note: $\boldsymbol{p}_{\boldsymbol{t}}$ for $\boldsymbol{t} \to \infty$ doesn't always converge!

Stationary Distribution

- Long-term average:

$$a_{\boldsymbol{t}} = \frac{1}{\boldsymbol{t}} \sum_{\boldsymbol{i}=0}^{\boldsymbol{t-1}} \boldsymbol{p_i}$$

- **Thm.** If G is strongly connected then $a_{\boldsymbol{t}} \to \boldsymbol{\pi}$:
  - $\boldsymbol{\pi} P = \boldsymbol{\pi}$
  - $\sum_i \boldsymbol{\pi}_i = 1$
  - $\boldsymbol{\pi}[P - I, \mathbf{1}] = [\mathbf{0}, 1]$
- We will show that $[P - I, \mathbf{1}]$ has rank $n \Rightarrow$ there is a unique solution to $\boldsymbol{\pi}[P - I, \mathbf{1}] = [\mathbf{0}, 1]$

Stationary Distribution Theorem

- **Thm.** $n \times (n+1)$ matrix $[P - I, \mathbf{1}]$ has rank n
- $A = [P - I, \mathbf{1}]$
- *Rank(A)* $< n \Rightarrow$ two lin. indep. solutions to *A**x**=0*
- $\sum_j p_{ij} = 1 \Rightarrow \sum_j p_{ij} - 1 = 0$ (row sums of A)
  - $(\mathbf{1}, 0)$ is a solution to *A**x** = 0*
- Assume there is another solution $(\boldsymbol{x}, \boldsymbol{\alpha}) \perp (\mathbf{1}, 0)$
  - $(P - I)\boldsymbol{x} + \boldsymbol{\alpha}\mathbf{1} = \mathbf{0}$
  - $\forall i$: $\sum_j p_{ij} x_j - x_i + \alpha = 0 \Rightarrow x_i = \sum_j p_{ij} x_j + \alpha$
- $(\boldsymbol{x}, \boldsymbol{\alpha}) \perp (\mathbf{1}, 0) \Rightarrow$ not all $\boldsymbol{x}_j$ are equal

Stationary Distribution Theorem Cont.

- $\forall i\colon\ x_i = \sum_j p_{ij} x_j + \boldsymbol{\alpha}$
- $(\boldsymbol{x}, \alpha) \perp (\mathbf{1}, 0) \Rightarrow$ not all $\boldsymbol{x}_j$ are equal
- $\boldsymbol{S} = \left\{i\colon x_i = Max_{j=1}^{n} x_j\right\} =$ set of max value coord.
  - $\bar{\boldsymbol{S}}$ is non-empty
- G strongly connected $\Rightarrow \exists\ edge\ (k, l)\colon k \in \boldsymbol{S}, l \in \bar{\boldsymbol{S}}$
- $\Rightarrow x_k > \sum_j p_{kj} x_j \Rightarrow \boldsymbol{\alpha} > 0$
- Symmetric argument with $\boldsymbol{S} = \left\{i\colon x_i = Min_{j=1}^{n} x_j\right\}$
- $\Rightarrow x_{k'} < \sum_j p_{k'j} x_j \Rightarrow \boldsymbol{\alpha} < 0$
- Contradiction so $(\mathbf{1}, 0)$ is the unique solution

Fundamental Theorem of Markov Chains

- **Thm.** If P is transition matrix of a strongly connected Markov Chain and $a_{\boldsymbol{t}} = \frac{1}{\boldsymbol{t}} \sum_{\boldsymbol{i}=0}^{\boldsymbol{t-1}} \boldsymbol{p}_{\boldsymbol{i}}$:
  - There exists a unique $\boldsymbol{\pi}$: $\boldsymbol{\pi} P = \boldsymbol{\pi}$
  - For any starting distribution: $\exists \lim_{\boldsymbol{t} \to \infty} a_{\boldsymbol{t}} = \boldsymbol{\pi}$
- $a_{\boldsymbol{t}}$ is a probability vector
- After one step: $a_{\boldsymbol{t}} \to a_{\boldsymbol{t}} P$
- $a_{\boldsymbol{t}} P - a_{\boldsymbol{t}} = \frac{1}{\boldsymbol{t}} \left[\sum_{i=0}^{\boldsymbol{t}-1} \boldsymbol{p}_i P\right] - \frac{1}{\boldsymbol{t}} \left[\sum_{i=0}^{\boldsymbol{t}-1} \boldsymbol{p}_i\right] = \frac{1}{\boldsymbol{t}} \left[\sum_{i=1}^{\boldsymbol{t}} \boldsymbol{p}_i\right] - \frac{1}{\boldsymbol{t}} \left[\sum_{i=0}^{\boldsymbol{t}-1} \boldsymbol{p}_i\right] = \frac{1}{\boldsymbol{t}} (\boldsymbol{p}_{\boldsymbol{t}} - \boldsymbol{p}_0)$
- $b_{\boldsymbol{t}} = a_{\boldsymbol{t}} P - a_{\boldsymbol{t}}$ satisfies $||b_{\boldsymbol{t}}||_1 \leq \frac{2}{\boldsymbol{t}} \to 0$

Fundamental Theorem of Markov Chains

- $n \times (n+1)\ matrix\ \boldsymbol{A} = [P - I, \mathbf{1}]$ has rank n
- $n \times n\ matrix\ \boldsymbol{B}$ = last n columns of $\boldsymbol{A}$
- First n columns of $\boldsymbol{A}$ sum to zero $\Rightarrow rank(\boldsymbol{B}) = n$
- $c_{\boldsymbol{t}}$ from $b_{\boldsymbol{t}} = a_{\boldsymbol{t}}P - a_{\boldsymbol{t}}$ by dropping first entry
- $a_{\boldsymbol{t}}B = [c_{\boldsymbol{t}}, 1] \Rightarrow a_{\boldsymbol{t}} = [c_{\boldsymbol{t}}, 1]B^{-1}$
- $b_{\boldsymbol{t}} \rightarrow 0 \Rightarrow [c_{\boldsymbol{t}}, 1] \rightarrow [\mathbf{0}, 1] \Rightarrow a_{\boldsymbol{t}} \rightarrow [\mathbf{0}, 1]B^{-1}$
- Let $[\mathbf{0}, 1]B^{-1} = \boldsymbol{\pi}$.
- Since $a_{\boldsymbol{t}} \rightarrow \boldsymbol{\pi}$ vector $\boldsymbol{\pi}$ is a probability distribution
- $a_{\boldsymbol{t}}[P - I] = b_{\boldsymbol{t}} = 0 \Rightarrow \boldsymbol{\pi}[P - I] = 0$