Forecasting

Lecturer: Prof. Duane S. Boning

Rev 8

Regression – Review & Extensions

- Single Model Coefficient: Linear Dependence $y = \beta x$
- Slope and Intercept (or Offset): $y = \beta_0 + \beta_1 x$
- Polynomial and Higher Order Models: $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- Multiple Parameters $y = \beta_0 + \beta_1 x + \beta_2 w$
- Key point: “linear” regression can be used as long as the model is linear in the coefficients (doesn’t matter the dependence in the independent variable)
- Time dependencies
  - Explicit $y = \beta_0 + \beta_1 t$
  - Implicit $y_i = \alpha \cdot y_{i-1} + w_i$

Agenda

1. Regression
   - Polynomial regression
   - Example (using Excel)

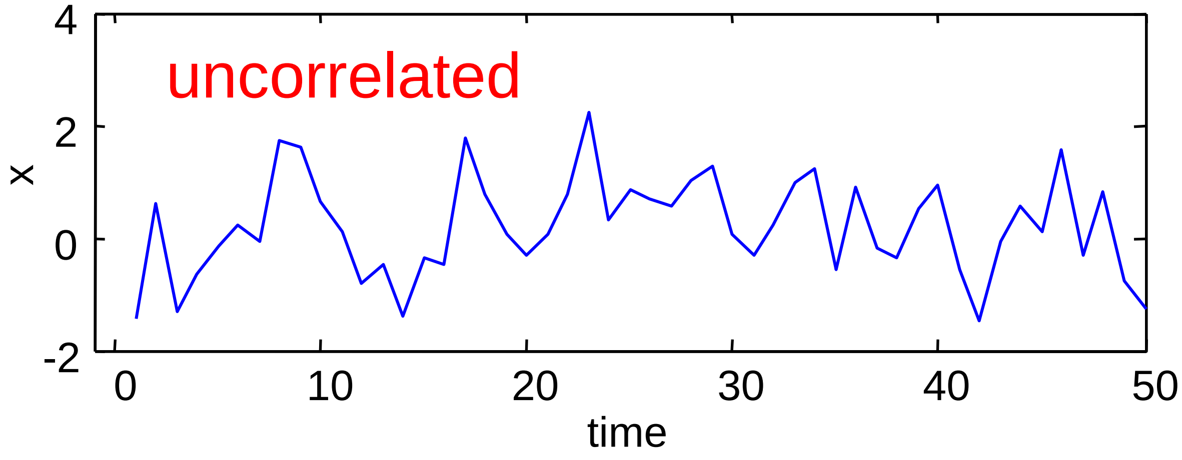
2. **Time Series Data & Time Series Regression**
   - **Autocorrelation – ACF**
   - **Example: white noise sequences**
   - **Example: autoregressive sequences**
   - **Example: moving average**
   - **ARIMA modeling and regression**

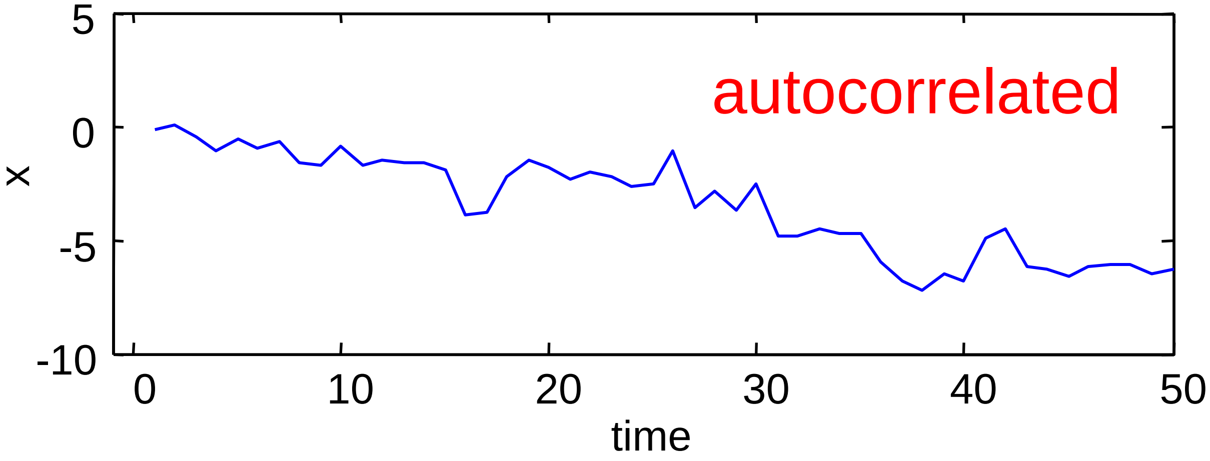
3. Forecasting Examples

Time Series – Time as an Implicit Parameter

- Data is often collected with a ***time-order***

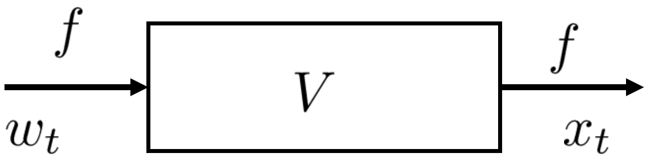
- An underlying dynamic process (e.g. due to physics of a manufacturing process) may create ***autocorrelation*** in the data





Intuition: Where Does Autocorrelation Come From?

- Consider a chamber with volume *V*, and with gas flow in and gas flow out at rate *f*. We are interested in the concentration *x* at the output, in relation to a known input concentration *w*.



$$\frac{dx_t}{dt} = (w_t - x_t)\frac{f}{V}$$

$$x_t = w_t - \frac{V}{f}\frac{dx_t}{dt} = w_t - \tau\frac{dx_t}{dt}$$

Consider a step change in input of w_0 at $t = 0$. Then

$$x_t = w_0(1 - e^{-t/\tau})$$

Discretizing: $$x_t = x_{t-1} + (w_0 - x_{t-1})(1 - e^{-\Delta t/T})$$

$$x_t = aw_t + (1-a)x_{t-1} \qquad \text{where } a = 1 - e^{-\Delta t/T}$$

correlation between x_t & x_{t-1} is $\rho = 1 - a = e^{-\Delta t/T}$

Key Tool: Autocorrelation Function (ACF)

- Time series data: time index *i*

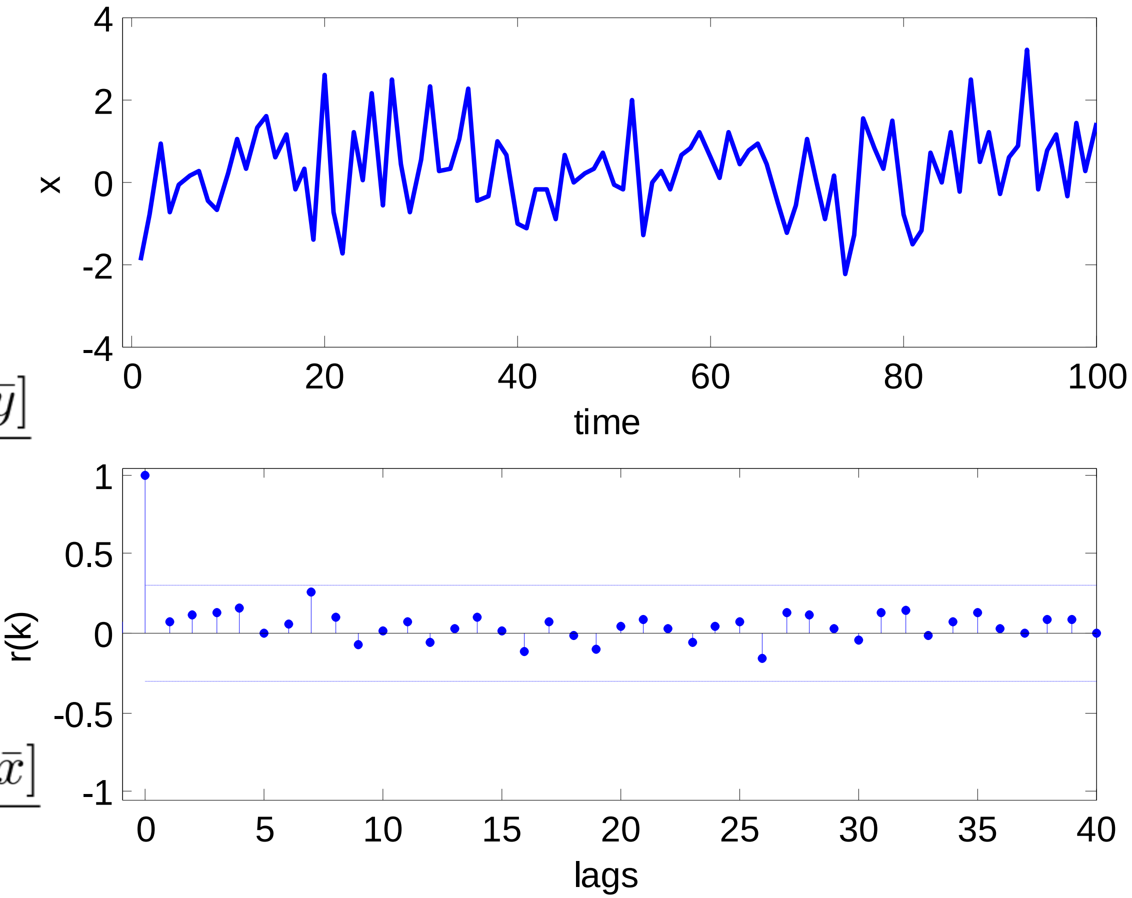
$$x_i \sim N(0,1)$$

- CCF: cross-correlation function

$$r_{xy}(k) = \frac{1}{N-1}\sum_{i=1}^{N-k}\frac{[x_i-\bar{x}][y_{i+k}-\bar{y}]}{s_x s_y}$$

- ACF: auto-correlation function

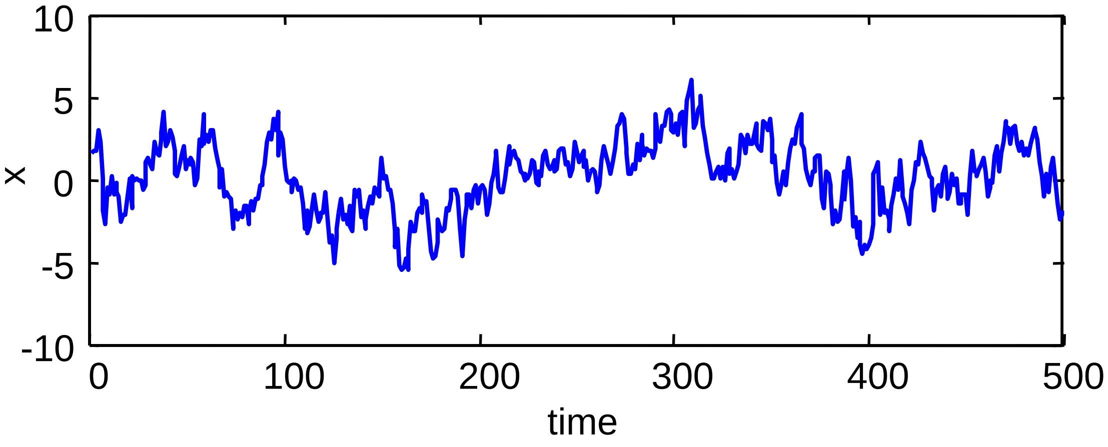
$$r_{xx}(k) = \frac{1}{N-1}\sum_{i=1}^{N-k}\frac{[x_i-\bar{x}][x_{i+k}-\bar{x}]}{s_x^2}$$

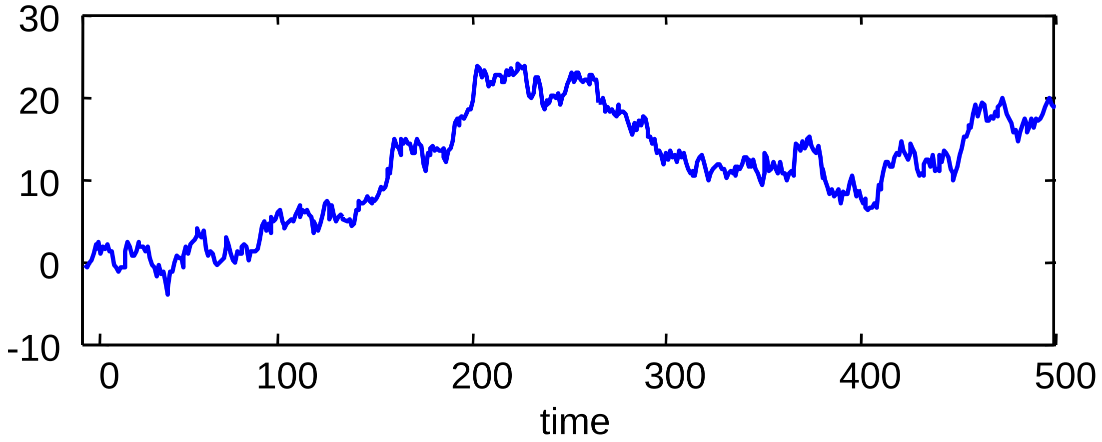


⇒ ACF shows the "similarity" of a signal to a lagged version of same signal

Stationary vs. Non-Stationary

Stationary series:
Process has a **fixed** mean





White Noise – An Uncorrelated Series

- Data drawn from IID gaussian

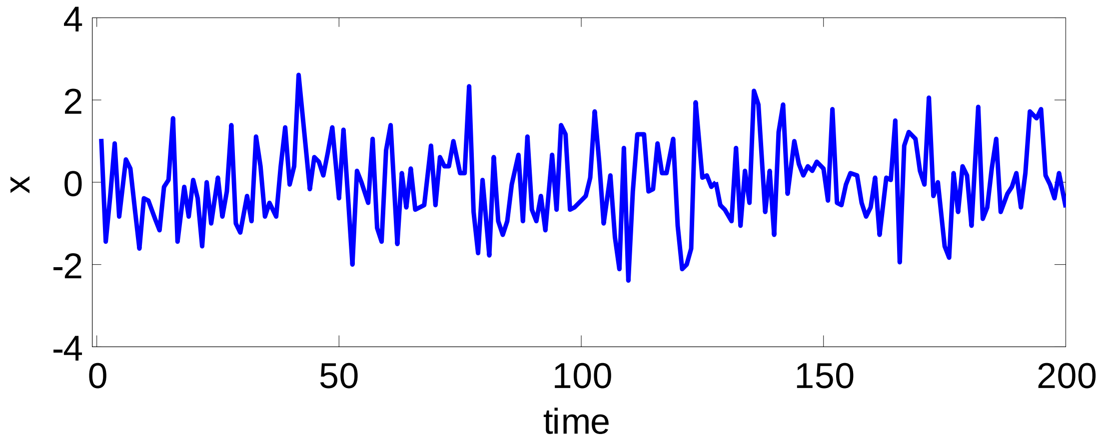
$$w_i \sim N(0,1)$$

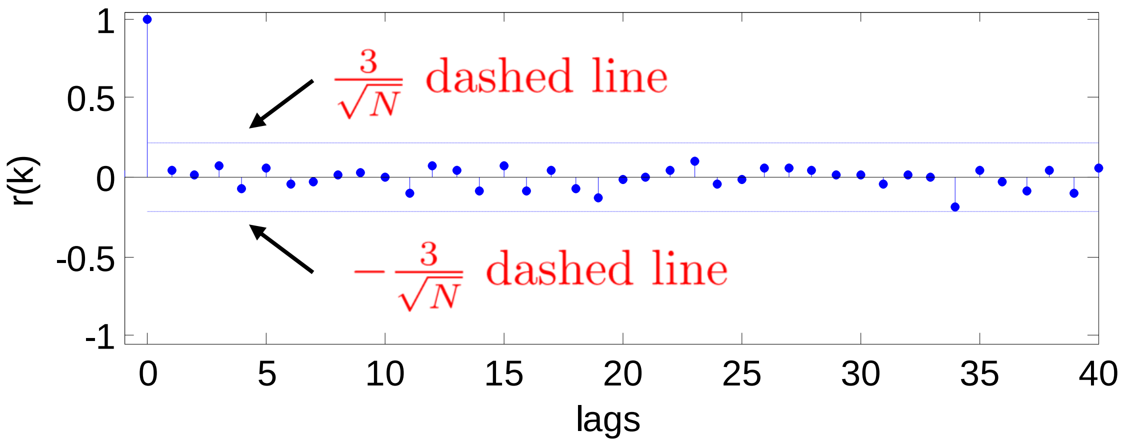
- ACF: We also plot the 3σ limits – values within these not significant
- Note that r(0) = 1 always (a signal is always equal to itself with zero lag – perfectly autocorrelated at k = 0)
- Sample mean

$$\bar{w} = \frac{1}{N}\sum_{i=1}^{N} w_i$$

- Sample variance

$$s_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w_i - \bar{w})^2$$





Autoregressive Disturbances

- Generated by:

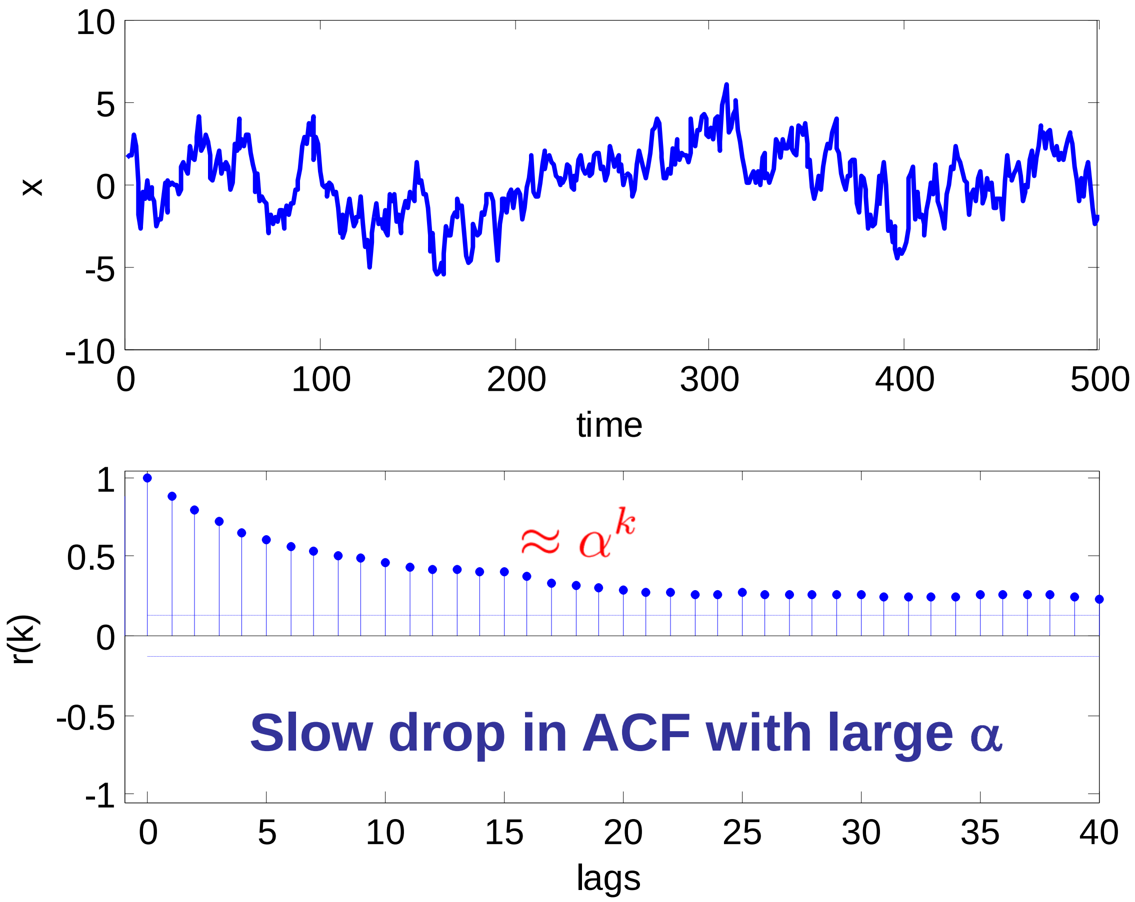
$$w_i \sim N(0,1)$$
$$c_i = \alpha \cdot c_{i-1} + w_i$$
$$\text{Shown: } \alpha = 0.9$$

- Mean

$$\mu_c = \mathrm{E}(c_i) = 0$$
$$\text{since } \mu_w = 0$$

- Variance

$$\begin{aligned} \sigma_c^2 &= \mathrm{Var}(c_i) = \mathrm{E}[(c_i - \bar{c})^2] \\ &= \mathrm{E}(\alpha^2 c_{i-1}^2 + 2\alpha c_{i-1} w_i + w_i^2) \\ &= \alpha^2 \mathrm{Var}(c_i) + \mathrm{Var}(w_i) \end{aligned}$$



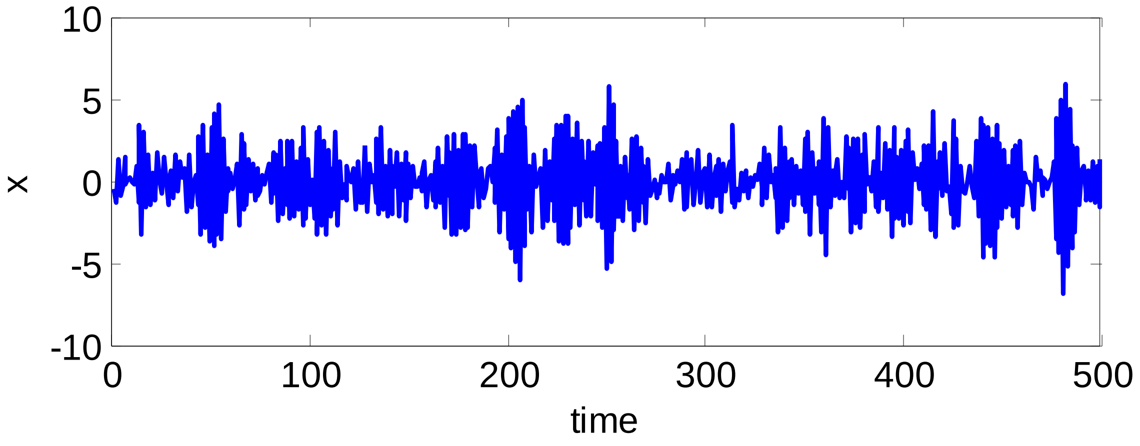
$$\Rightarrow \quad \sigma_c^2 = \frac{\sigma_w^2}{1 - \alpha^2}$$

So AR (autoregressive) behavior *increases* variance of signal.

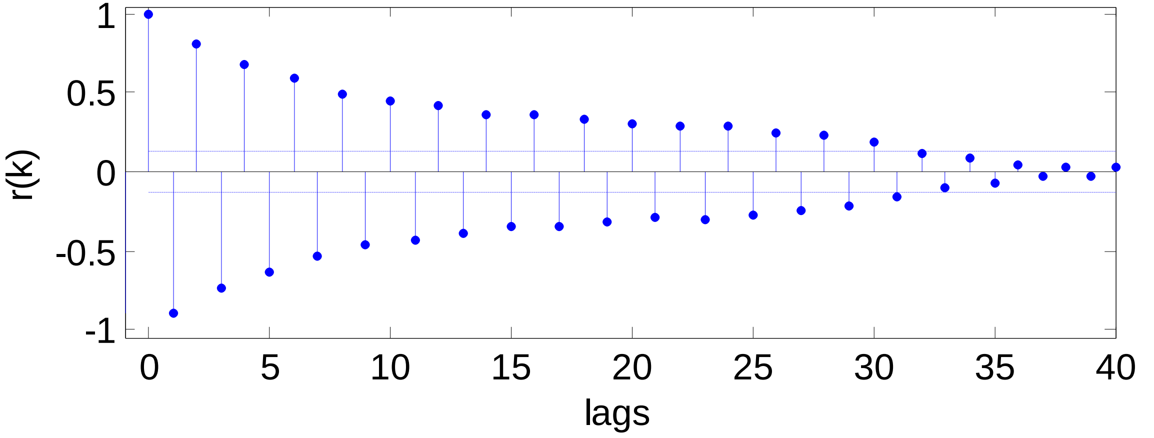
Another Autoregressive Series

- Generated by:

$$w_i \sim N(0,1)$$
$$c_i = \alpha \cdot c_{i-1} + w_i$$
$$\text{Shown: } \alpha = -0.9$$



- High **negative** autocorrelation:



Slow drop in ACF with large α

But now ACF alternates in sign

Random Walk Disturbances

- Generated by:

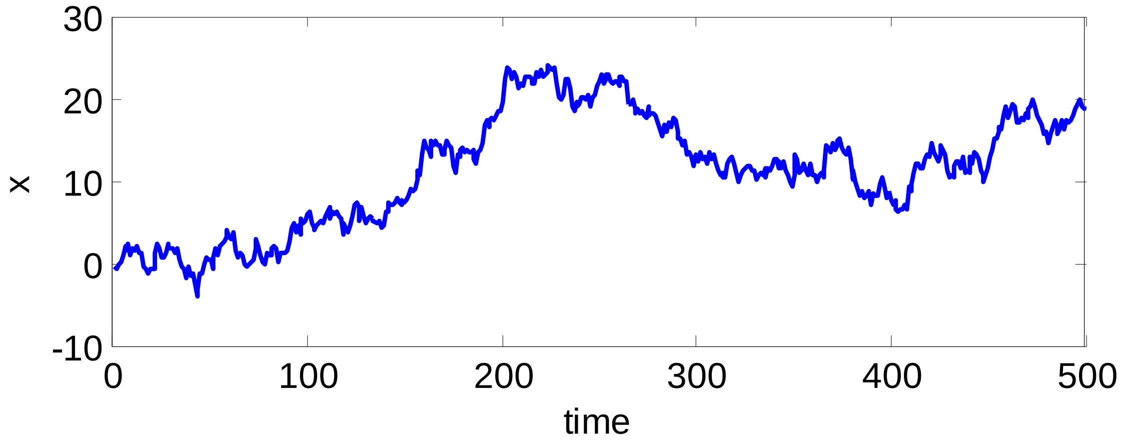
$$w_i \sim N(0,1)$$
$$c_i = 1 \cdot c_{i-1} + w_i$$
$$\text{AR with } \alpha = 1$$

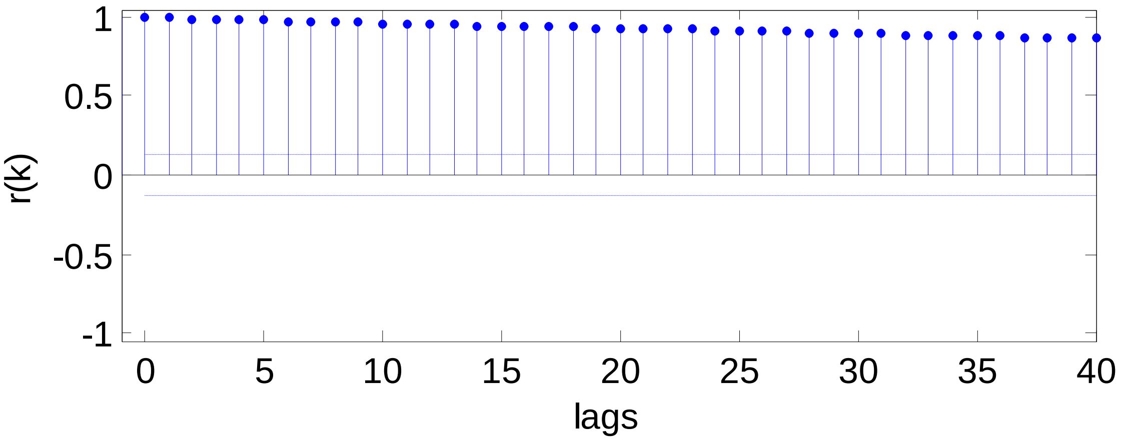
- Mean

$$\bar{c} \neq 0 \quad \text{non-stationary}$$

- Variance

Variance increases as sequence gets longer





Very **slow drop in ACF for $\alpha = 1$**

Moving Average Sequence

- Generated by:

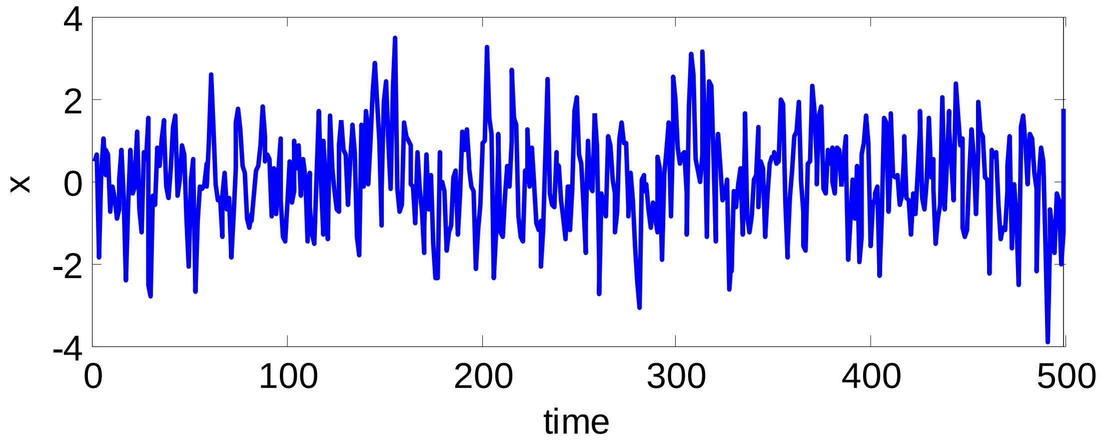
$$w_i \sim N(0,1)$$
$$c_i = w_i + \beta \cdot w_{i-1}$$
$$\text{Shown: } \beta = 0.5$$

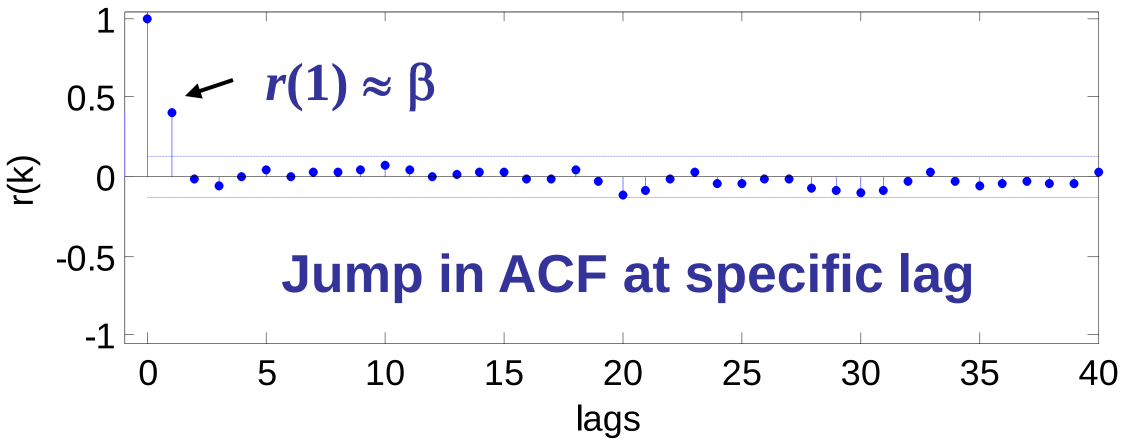
- Mean

$$\mu_c = \mathrm{E}(c_i) = 0$$
$$\text{since } \mu_w = 0$$

- Variance

$$\begin{aligned}\sigma_c^2 &= \mathrm{Var}(c_i) = \mathrm{E}[(c_i - \bar{c})^2] \\ &= \mathrm{E}(w_i^2 + 2\beta w_i w_{i-1} + \beta^2 w_{i-1}^2) \\ &= (1+\beta^2)\mathrm{Var}(w_i)\end{aligned}$$





$$\Rightarrow \quad \sigma_c^2 = (1+\beta^2)\sigma_w^2$$

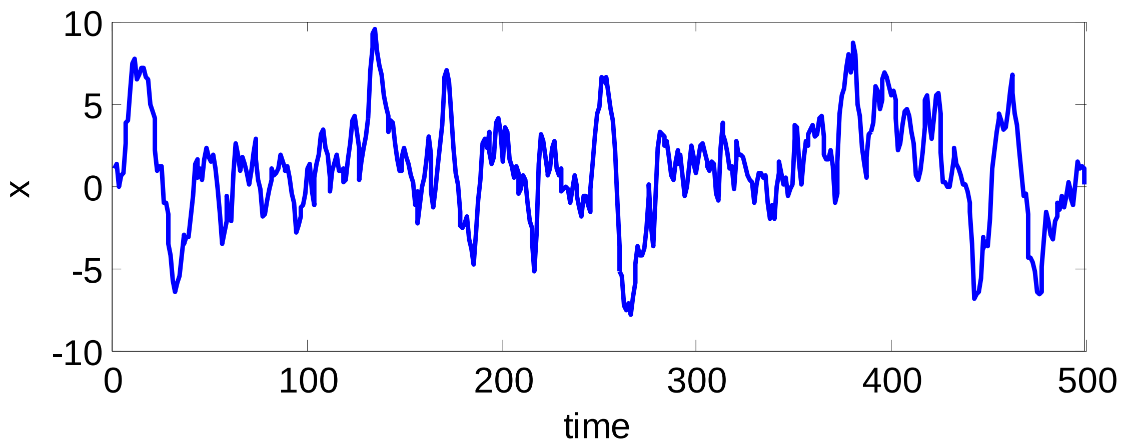
So MA (moving average) behavior also *increases* variance of signal.

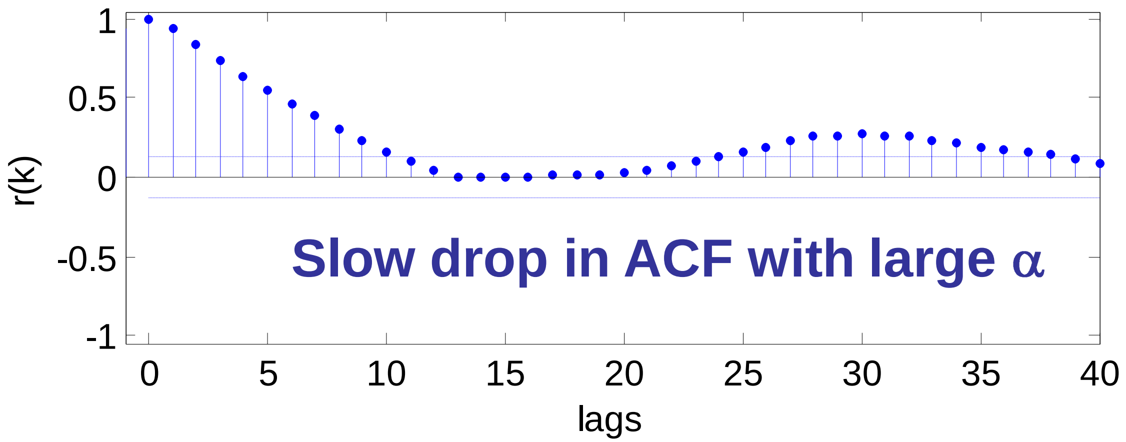
ARMA Sequence

- Generated by:

$$w_i \sim N(0,1)$$
$$c_i = \alpha \cdot c_{i-1} + w_i + \beta \cdot w_{i-1}$$
$$\text{Shown: } \alpha = 0.9, \quad \beta = 0.5$$

- Both AR & MA behavior





ARIMA Sequence

- Start with ARMA sequence:

$$w_i \sim N(0,1)$$

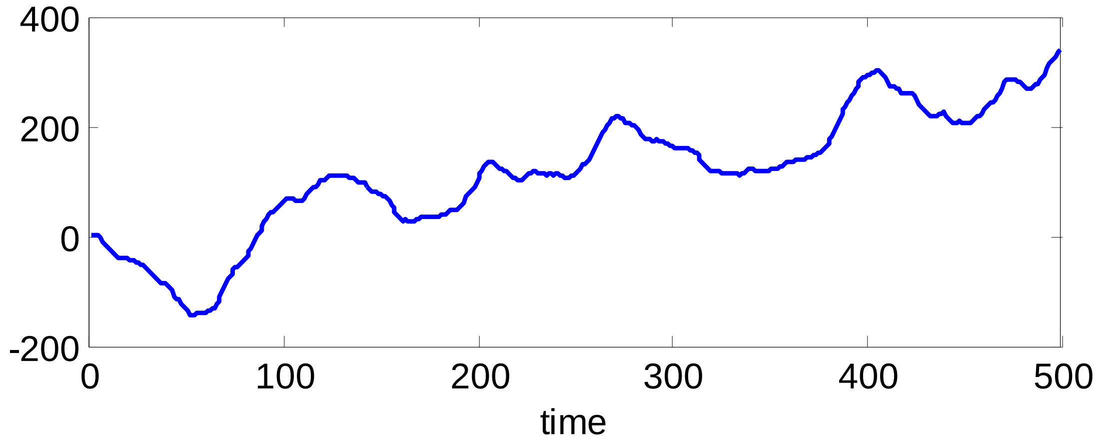
$$c_i = \alpha \cdot c_{i-1} + w_i + \beta \cdot w_{i-1}$$

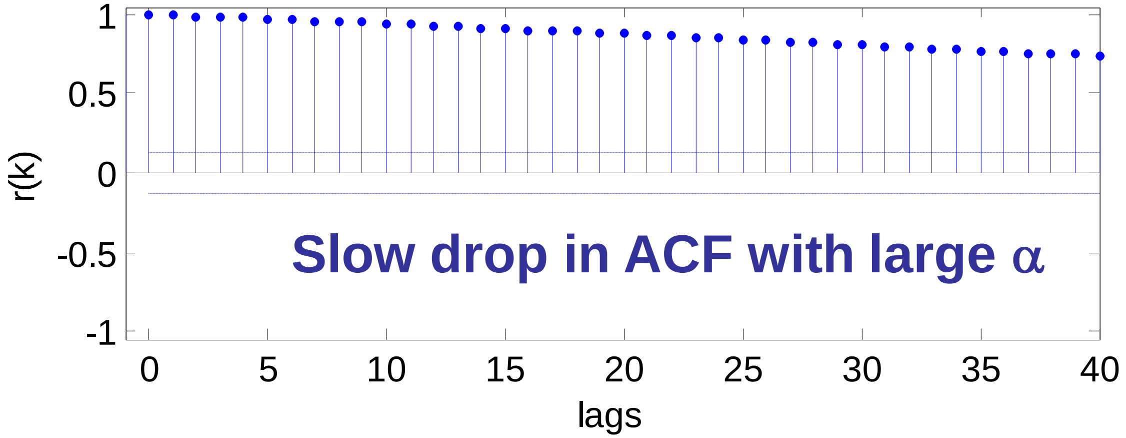
$$\text{Shown: } \alpha = 0.9, \quad \beta = 0.5$$

- Add Integrated (I) behavior

$$x_i = x_{i-1} + c_i$$

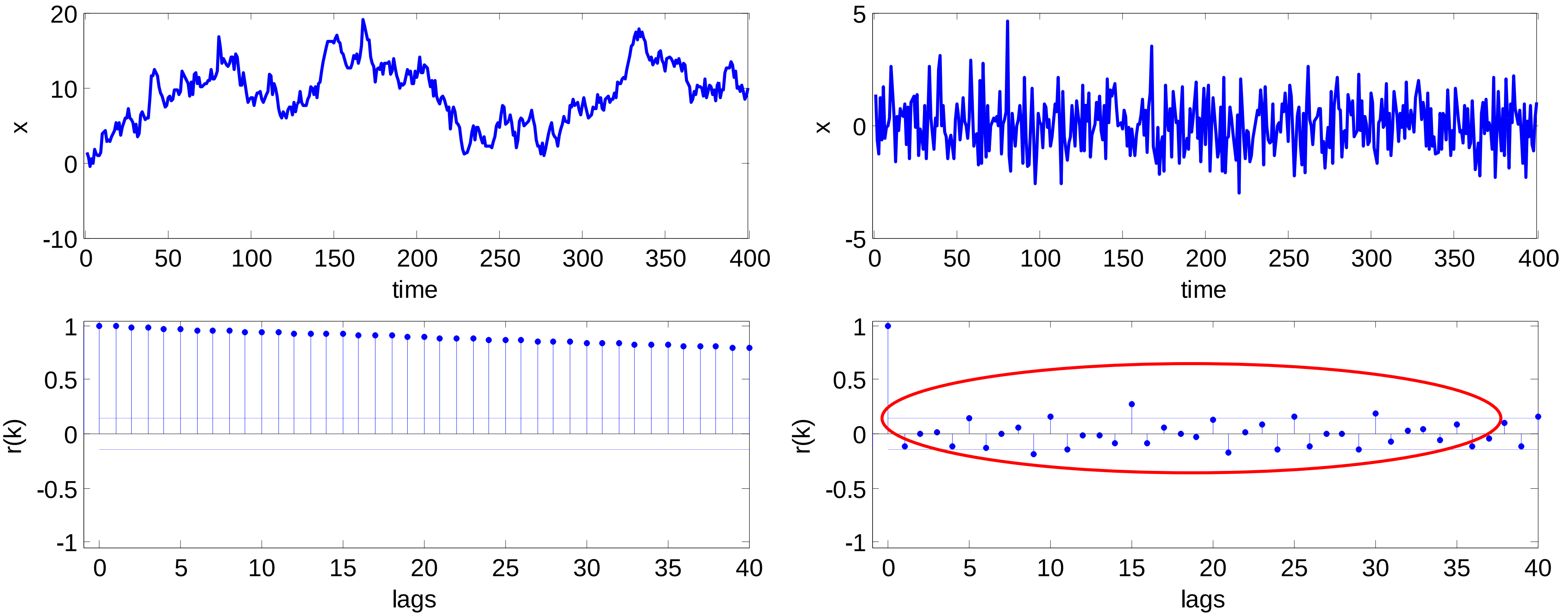
random walk (integrative) action





Periodic Signal with Autoregressive Noise

Original Signal | After Differencing



$$d_i = x_i - x_{i-1}$$

See underlying signal with period = 5

Cross-Correlation: A Leading Indicator

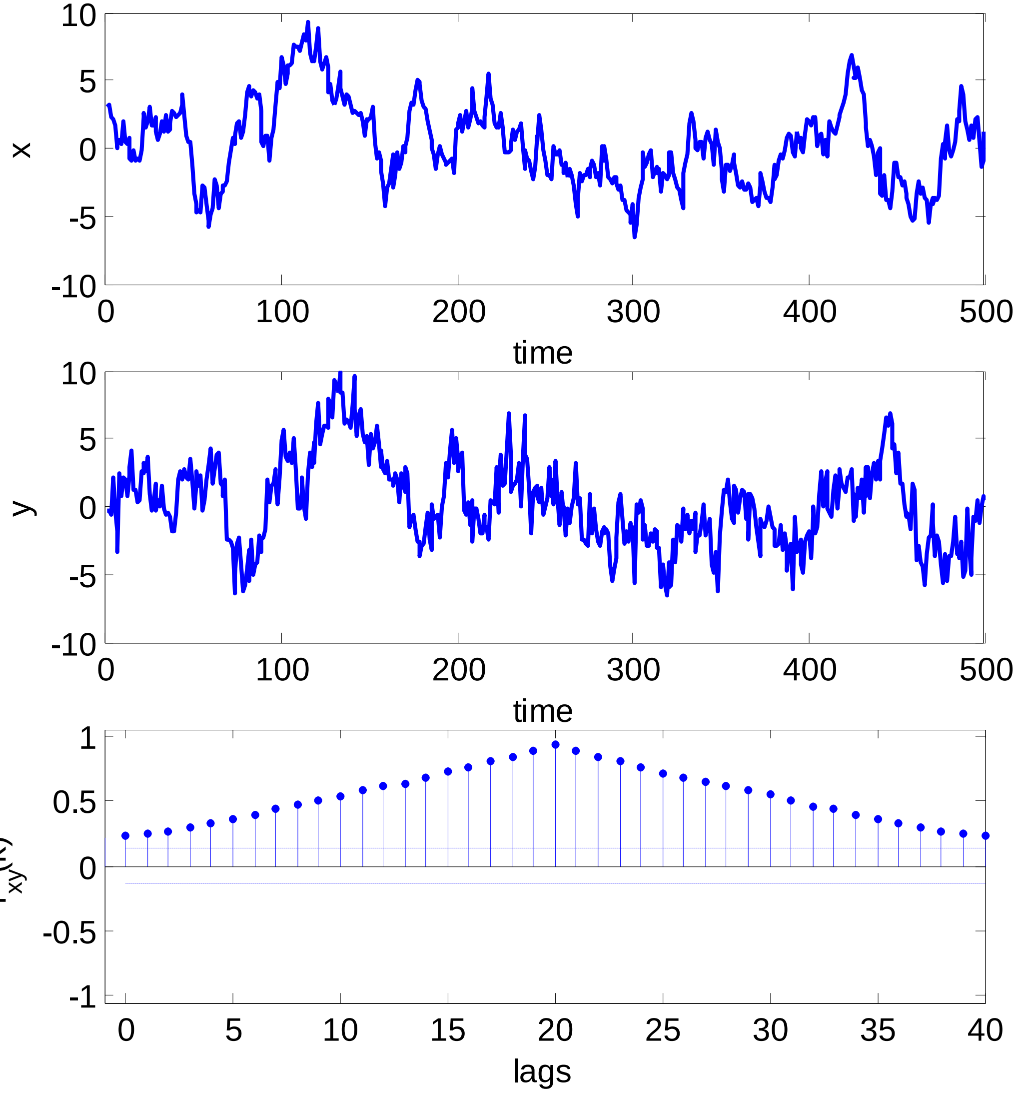
- Now we have two series:
  - An “input” or explanatory variable x
  - An “output” variable y

$$y_i = x_{i-k} + w_i$$

$$w_i \sim N(0,1)$$

Shown: lag $k = 20$ and autoregressive x with $\alpha = 0.9$

- CCF indicates both AR and lag:



Regression & Time Series Modeling

- The ACF or CCF are helpful tools in selecting an appropriate model structure
  - Autoregressive terms?
    - $x_i = \alpha \, x_{i-1}$
  - Lag terms?
    - $y_i = \gamma \, x_{i-k}$
- One can structure data and perform regressions
  - Estimate *model coefficient* values, significance, and confidence intervals
  - Determine confidence intervals on *output*
  - Check residuals

Statistical Modeling Summary

1. Statistical Fundamentals
   - Sampling distributions
   - Point and interval estimation
   - Hypothesis testing

2. Regression
   - ANOVA
   - Nominal data: modeling of treatment effects (mean differences)
   - Continuous data: least square regression $y = f(\mathbf{x}, \mathbf{b})$

3. Time Series Data & Forecasting
   - Autoregressive, moving average, and integrative behavior
   - Auto- and Cross-correlation functions $x_i = f(\mathbf{x}, \mathbf{b})$
   - Regression and time-series modeling $y_i = f(\mathbf{x}, \mathbf{b})$

2.854 / 2.853 Introduction to Manufacturing Systems
Fall 2010