ECON5345 Lecture 1: Statistics Review

Byoungchan Lee

February 3, 2026

1 Basic Asymptotics

1.1 Law of Large Numbers and Central Limit Theorem

For a covariance-stationary stochastic process $\{x_t\}$, under some regularity conditions,

$$\text{LLN:} \quad \frac{1}{T}\sum_{t=1}^{T} x_t \xrightarrow{p/a.s.} \mu_x,$$

$$\text{CLT:} \quad \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} x_t - \mu_x\right) \xrightarrow{d} N(0, \text{LRV}),$$

where the long-run variance

$$\text{LRV} = \cdots + \Gamma_{-1} + \Gamma_0 + \Gamma_1 + \cdots \quad \text{and} \quad \Gamma_j = \text{cov}(x_t, x_{t-j}).$$

1.2 Some asymptotic theory

Let g be a continuous/differentiable function (e.g., a mapping from VAR coefficients to impulse response functions / forecast error variance decompositions).

Continuous Mapping Theorem

If

$$X_T \xrightarrow{p/d/a.s.} X,$$

then

$$g(X_T) \xrightarrow{p/d/a.s.} g(X).$$

ex)

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V) \;\Rightarrow\; \sqrt{T}(\hat{\theta}_T - \theta_0)' V^{-1} \sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \|N(0, I)\|^2 = \chi^2(\text{Rank}(V)).$$

1.3 Slutsky's Theorem

If $X_T \xrightarrow{d} X$ and $Y_T \xrightarrow{p} y$ (constant), then

$$X_T + Y_T \xrightarrow{d} X + y,$$
$$X_T Y_T \xrightarrow{d} Xy,$$
$$X_T / Y_T \xrightarrow{d} X/y \quad \text{if } y \neq 0.$$

ex)

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V), \quad \hat{V} \xrightarrow{p} V \;\Rightarrow\; \sqrt{T}\hat{V}^{-1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, I),$$
$$\sqrt{T}(\hat{\theta}_T - \theta_0)'\hat{V}^{-1}\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \|N(0, I)\|^2 = \chi^2(\text{Rank}(V)).$$

1.4 Delta Method

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V) \;\Rightarrow\; \sqrt{T}\big(g(\hat{\theta}_T) - g(\theta_0)\big) \xrightarrow{d} N(0, G'VG),$$

where

$$G' = \left.\frac{\partial g}{\partial \theta'}\right|_{\theta=\theta_0}.$$

ex) Consider an AR(1) process

$$x_t = \rho x_{t-1} + e_t, \qquad e_t \sim i.i.d.\ N(0, \sigma^2).$$

Given $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{d} N(0, V)$, what can we say about the IRF $\{1, \rho, \rho^2, \ldots\}$?

ans)

$$\sqrt{T}\big([\hat{\rho}]^h - \rho^h\big) \xrightarrow{d} N\Big(0,\ (h[\hat{\rho}]^{h-1})^2 V\Big).$$

2 Estimators and their Economic Applications

2.1 OLS

Model, estimation, and identification

Consider the linear regression model

$$y_t = x_t'\beta + e_t,$$

and define the OLS estimator as the minimizer of the sum of squared residuals,

$$\min_{\beta} \sum_t (y_t - x_t'\beta)^2.$$

The identification assumption is $E[x_t e_t] = 0$ and $\Sigma_{xx} = E[x_t x_t']$ is non-singular. Threats to identification include measurement errors in x_t (it usually matters when x_t is a generated regressor; if the null hypothesis is $\beta = 0$, we do not need to adjust inferences although there exist measurement errors (Pagan, 1984)), omitted variable bias, and reverse-causality.

Estimator, consistency, and asymptotic normality

The estimator is

$$\begin{aligned}\hat{\beta} &= \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \left(\sum_{t=1}^{T} x_t y_t\right) \\ &= \beta + \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} x_t e_t\right).\end{aligned}$$

Consistency follows as

$$\hat{\beta} \xrightarrow{p} \beta + \Sigma_{xx}^{-1} 0 = \beta.$$

For asymptotic normality,

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t e_t\right) \xrightarrow{d} \Sigma_{xx}^{-1} N\big(0, \mathrm{LRV}(x_t e_t)\big).$$

For statistical inference, we need to consistently estimate $\mathrm{LRV}(x_t e_t)$.

2.2 Coibion and Gorodnichenko (2015, AER)

Specification and objects

$$\pi_{t+3,t} - F_t \pi_{t+3,t} = c + \beta\big(F_t \pi_{t+3,t} - F_{t-1}\pi_{t+3,t}\big) + \mathrm{error}_t.$$

Here, $\pi_{t+3,t}$ is the average inflation rate over the current and next three quarters, and F_t is the average time-t forecast across agents.

Theoretical predictions

Under FIRE, $\beta = 0$. Under sticky information (Mankiw and Reis, 2002),

$$\beta = \frac{\lambda}{1-\lambda},$$

where λ is the probability of no information update.

Reported estimate

$$\beta \approx 1, \qquad \lambda \approx 0.5, \qquad \text{Information is updated every 2 quarters on avg.}$$

2.3 IV

Model, assumptions, and estimator

Consider

$$y_t = x_t'\beta + e_t, \qquad \text{(stacked): } Y = X\beta + e.$$

The identification assumption is $E[z_t e_t] = 0$, $\Sigma_{zx} = E[z_t x_t']$ has a full rank, and $\Sigma_{zz} = E[z_t z_t']$ is non-singular. The IV estimator is

$$\hat{\beta}_{IV} = (X'P_Z X)^{-1}(X'P_Z Y) = \beta + (X'P_Z X)^{-1}(X'P_Z e),$$

where $P_Z = Z(Z'Z)^{-1}Z$. Note that

$$\frac{1}{T}X'Z = \frac{1}{T}\sum_{t=1}^{T} x_t z_t' \xrightarrow{p} E[x_t z_t'].$$

Consistency, asymptotic normality, and weak instruments

Consistency:

$$\hat{\beta}_{IV} \xrightarrow{p} \beta + (\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})^{-1}(\Sigma_{xz}\Sigma_{zz}^{-1}E[z_t e_t]) = \beta.$$

Asymptotic Normality:

$$\begin{aligned}\sqrt{T}(\hat{\beta}_{IV} - \beta) &\approx (\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})^{-1}\left(\Sigma_{xz}\Sigma_{zz}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t e_t\right) \\ &\xrightarrow{d} (\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})^{-1}(\Sigma_{xz}\Sigma_{zz}^{-1})N\big(0, \mathrm{LRV}(z_t e_t)\big).\end{aligned}$$

For statistical inference, we need to consistently estimate $\mathrm{LRV}(z_t e_t)$. Finding a proper IV is not easy.

Weak IV: what if $P_Z X \approx 0$? Then

$$\hat{\beta}_{IV} = (X'P_Z X)^{-1}(X'P_Z Y) \approx 0^{-1}(X'P_Z Y),$$

so $\hat{\beta}_{IV}$ can easily blow up and its sign can be easily flipped. The sampling distribution of $\hat{\beta}_{IV}$ may be very different from the asymptotic normal distribution above. (First-stage.) We can directly check whether $P_Z X \approx 0$ or not. What if your IV is weak? See Andrews, Stock and Sun (2019).

2.4 Coibion and Gorodnichenko (2012, AEJ: Macro)

Persistence and interpretation

The target FFR i_t is very persistent. Two explanations for the persistence are interest rate smoothing ($\rho_{i,k} > 0$) and persistent monetary policy shocks ($\rho_{u,j} > 0$). How to test the two hypotheses? If $\rho_{u,j} > 0$ is the primary reason for the persistence in i_t, the response of i_t to nonmonetary policy shocks should be less persistent. Use nonmonetary shocks as IVs.

Reported estimate

$$\rho_i \gg 0, \qquad \rho_u \approx 0.$$

2.5 Nakamura and Steinsson (2014, AER)

Regression and instrument

Want to estimate G military spending multiplier β:

$$\frac{Y_{i,t} - Y_{i,t-2}}{Y_{i,t-2}} = \alpha_i + \gamma_t + \beta \frac{G_{i,t} - G_{i,t-2}}{Y_{i,t-2}} + \varepsilon_{it},$$

where Y_{it} and G_{it} are per capital output and military spending in state i in year t. There exists an obvious endogeneity issue here. (why?) The instrument is $\left\{ D_i \frac{G_t - G_{t-2}}{Y_{t-2}} \right\}$. Justification: (1) national military spending G_t is mostly driven by geopolitical events, and (2) the sensitivity of G_{it} to G_t varies across states. Identifying assumption: "the United States does not embark on a military buildup because states that receive a disproportionate amount of military spending are doing poorly relative to other states."

Multiplier comparison

Open economy relative multiplier > closed-econ multiplier: the responses of monetary policy stance and aggregate taxation are differenced out by the time-fixed effects.

2.6 MoM

Moment conditions and examples

Moment conditions: there uniquely exists θ_0 such that

$$g(\theta_0) = 0,$$

where $g(\theta) = E[h(\theta, x_t)]$, and g and θ are $r \times 1$ vectors. The model is (just identified) when # of parameters = # of moment conditions. Let

$$g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h(\theta, x_t), \qquad g_T(\hat{\theta}_T) = 0.$$

Example:

$$h(\mu, \sigma^2, x) = \begin{pmatrix} x - \mu \\ \sigma^2 - (x-\mu)^2 \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 \end{pmatrix}.$$

Also,

$$h(\beta, x_t, y_t) = x_t y_t - x_t x_t' \beta \Rightarrow \text{OLS}.$$

From conditional moments to unconditional moments

Consider the Euler equation,

$$E_t\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(1 + r_{t+1}) - 1\right] = 0.$$

By the L.I.E.,

$$E\left[\left\{\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(1 + r_{t+1}) - 1\right\} z_t\right] = 0 \quad \text{for any } z_t \in \mathcal{F}_t.$$

We may consider $z_t = (1,\, C_t/C_{t-1})'$ to estimate β and γ. Moments in g may include impulse response coefficients, forecast error variance decompositions, mean, variance, auto-covariances, cross-correlations, cross-sectional distributions, etc.

Heuristic derivation of the asymptotic distribution

$$0 = g_T(\hat{\theta}_T) \approx g_T(\theta_0) + G_T'(\hat{\theta}_T - \theta_0),$$

where $G_T' = \left.\frac{\partial g_T}{\partial \theta'}\right|_{\theta_0}$. Then,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \approx -(G')^{-1}\sqrt{T} g_T(\theta_0),$$

where

$$G_T' = \left.\frac{\partial g_T}{\partial \theta'}\right|_{\theta_0} \xrightarrow{p} \left.\frac{\partial g}{\partial \theta'}\right|_{\theta_0} = G',$$

which is assumed to be invertible. Note that

$$\sqrt{T} g_T(\theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(\theta_0, x_t) \xrightarrow{d} N\big(0, \text{LRV}\big(h(\theta_0, x_t)\big)\big),$$

$$\therefore\ \sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\big(0,\ (G')^{-1}\text{LRV}\big(h(\theta_0, x_t)\big)(G)^{-1}\big).$$

2.7 Kocherlakota (1996, JEL)

Equity premium puzzle based on aggregate consumption

Equity Premium Puzzle based on Aggregate Consumption.

$$E_t\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(1+r^i_{t+1})\right]=1 \qquad \text{for } i=\text{bond, stock.}$$

Take the difference, divide by β, and apply the L.I.E.:

$$E\left[\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(r^s_{t+1}-r^b_{t+1})\right]=0.$$

When

$$e_t=\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(r^s_{t+1}-r^b_{t+1}),$$

$\bar{e}$ should be close to zero.

Slide figure

[Slide content appears as a figure in the original PDF.]

2.8 Toda and Walsh (2015, JPE)

Individual Euler equation and a MoM estimator

Sometimes, LLN and CLT may not work!

Consumption Euler equation at the individual level

$$1=E_{i,t-1}\left[\beta(c_{i,t}/c_{i,t-1})^{-\gamma}(1+r^j_t)\right] \quad \forall i \text{ and } j.$$

LIE

$$\Rightarrow\ 1=E\left[\beta\int\left(\frac{c_{i,t}}{c_{i,t-1}}\right)^{-\gamma}di\ (1+r^j_t)\right]$$

$$\Rightarrow\ 0=E\left[\int\left(\frac{c_{i,t}}{c_{i,t-1}}\right)^{-\gamma}di\ (r^{stock}_t-r^{bond}_t)\right].$$

Thus, we can consider a MoM estimator of γ by solving:

$$0=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{I}\sum_{i=1}^{I}\left(\frac{c_{i,t}}{c_{i,t-1}}\right)^{-\gamma}(r^{stock}_t-r^{bond}_t).$$

What if the cross-sectional moment is infinite?

$$0 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{I}\sum_{i=1}^{I}\left(\frac{c_{i,t}}{c_{i,t-1}}\right)^{-\gamma}(r_t^{stock} - r_t^{bond}).$$

However, what if

$$\int\left(\frac{c_{i,t}}{c_{i,t-1}}\right)^{-\gamma}di = \infty$$

for the true value of γ? Toda and Walsh show that the cross-sectional distribution of consumption growth features fat upper and lower tails. In this case, the lower tail may make $\int(c_{i,t}/c_{i,t-1})^{-\gamma}di = \infty$ for reasonable values of γ. As a result, $\hat{\gamma} \approx 0$.

2.9 GMM

Moment conditions and estimator

Moment conditions: there uniquely exists θ_0 such that $g(\theta_0) = 0$, where $g(\theta) = E[h(\theta, x_t)]$, and $\dim(g) = r \geq \dim(\theta) = a$. The model is (over identified) if $r > a$. Let $g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}h(\theta, x_t)$. The GMM estimator solves

$$\min_{\theta}\ g_T(\theta)'W_T g_T(\theta),$$

where $\{W_T\}$ is a sequence of p.d. weighting matrices such that $W_T \xrightarrow{p} W$. The first-order condition is

$$\left[\frac{\partial g_T(\hat{\theta}_T)}{\partial\theta'}\right]' W_T g_T(\hat{\theta}_T) = 0.$$

Heuristic derivation of the asymptotic distribution and optimal weighting

$$0 = \left[\frac{\partial g_T(\hat{\theta}_T)}{\partial\theta'}\right]' W_T g_T(\hat{\theta}_T) \approx \left[\frac{\partial g_T(\hat{\theta}_T)}{\partial\theta'}\right]' W_T\left[g_T(\theta_0) + G_T'(\hat{\theta}_T - \theta_0)\right],$$

where $G_T' = \frac{\partial g_T}{\partial\theta'}\Big|_{\theta_0}$. Then, for $\Omega = \mathrm{LRV}(h(\theta_0, x_t))$,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \approx -(GWG')^{-1}GW\sqrt{T}g_T(\theta_0) \Rightarrow \sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} (GWG')^{-1}GW\,N(0, \Omega).$$

The optimal weighting matrix is $W = \Omega^{-1}$. Then,

$$\mathrm{var}\big(\sqrt{T}(\hat{\theta}_T - \theta_0)\big) = (G\Omega^{-1}G')^{-1}.$$

Two-step (and iterated) GMM; overidentification tests

How to estimate the optimal weighting matrix? In a two-step approach, $\hat{\theta}_T$ is consistent for any W_T, so start with $W_T = I$. Estimate $\Omega = \text{LRV}(h(\theta_0, x_t))$ using $\{h(\hat{\theta}_T, x_t)\}$; denote the estimate by $\hat{\Omega}$. Set $W_T = \hat{\Omega}^{-1}$ (a consistent estimate of Ω^{-1}) and run GMM again. (Iterated GMM.) Iterate until $\hat{W}$ converges.

For overidentification tests (Hansen's J-test),

$$\sqrt{T} g_T(\theta_0) \xrightarrow{d} N(0, \Omega) \Rightarrow \sqrt{T} g_T(\theta_0)' \Omega^{-1} \sqrt{T} g_T(\theta_0) \xrightarrow{d} \chi^2(r), \quad r = \dim(g).$$

Based on consistent estimators $\hat{\theta}_T$ and $\hat{\Omega}$,

$$J = \sqrt{T} g_T(\hat{\theta}_T)' \hat{\Omega}^{-1} \sqrt{T} g_T(\hat{\theta}_T) \xrightarrow{d} \chi^2(r - a),$$

where $r = \dim(g)$ and $a = \dim(\theta)$. In theory, we can use this test to check whether our moment conditions are consistent with the data.

Practical notes

IV exogeneity can be tested. In practice, it is well-known that the finite-sample performances of both the optimal GMM and J-tests are not so good. It is advisable to supplement the results by trying several specifications when you use GMM.

2.10 Gourinchas and Parker (2002, ECTA)

Method of simulated moments / simulated moments

Using a Method of Simulated Moments (or SMM) to estimate a life-cycle model: a model of a typical household working from age 25 to 65 and retire thereafter, with uninsurable idiosyncratic earnings risk. MSM: pick parameter values, solve the model, simulate the moments of interest (20,000 income processes over 40 years in the paper), compare the moments based on the simulated data to the empirical counterpart, and iterate until the difference is minimized.

Reported fit

Given the estimated income process, $\beta = 0.96$ and RRA $= 0.514$ replicate the life-cycle pattern in consumption, derived from CEX, pretty well. When the optimal weighting matrix is used, RRA $= 1.4$.

2.11 MLE

Limited vs. full information; motivation

When the previous methods focus on a subset of the population properties (i.e. moments), they are considered as the limited information approach. In contrast, MLE is a full information approach in the sense that it requires us to specify the whole data-generating process, and we use the model structure in estimation. For example, suppose that the economy is driven by two shocks (say, a productivity shock and a monetary policy shock). If we estimate the model by matching the impulse responses to only monetary policy shocks, it is a limited information approach: we do not need to fully specify how the productivity shocks propagate in the model. In contrast, to use MLE, we need a fully specified DGP. Because we use the "full" information, it is (asymptotically) efficient when the model is correctly specified.

Likelihood, asymptotics, and information matrix estimation

Let $X^T = \{x_T, x_{T-1}, \ldots, x_1\}$ and suppose the joint density is $f(X^T \mid \theta)$. The likelihood is $L(\theta \mid X^T) \equiv f(X^T \mid \theta)$ and the log-likelihood is $\ell(\theta \mid X^T) = \log(L(\theta \mid X^T))$. Estimation: maximize $L(\theta \mid X^T)$ or $\ell(\theta \mid X^T)$.

Asymptotically,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, I^{-1}),$$

where I is the Fisher information matrix. Two methods of estimating the information matrix:

$$I = -E\left[\frac{\partial^2 \ell(\theta_0)/T}{\partial\theta\,\partial\theta'}\right] \Rightarrow \hat{I}_1 = -\frac{1}{T}\frac{\partial^2 \ell(\hat{\theta}_T)}{\partial\theta\,\partial\theta'},$$

$$I = E\left[\frac{\partial \log f(x_t \mid x_{t-1}, \theta_0)}{\partial\theta} \cdot \frac{\partial \log f(x_t \mid x_{t-1}, \theta_0)}{\partial\theta'}\right],$$

$$\hat{I}_2 = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial \log f(x_t \mid x_{t-1}, \hat{\theta}_T)}{\partial\theta} \cdot \frac{\partial \log f(x_t \mid x_{t-1}, \hat{\theta}_T)}{\partial\theta'}\right].$$

Model misspecification and quasi-maximum likelihood

If the model is correctly specified, $\hat{I}_1$ and $\hat{I}_2$ should be similar; if not, the model might be misspecified. White (1982) proposes the "quasi-maximum likelihood" standard error that is valid sometimes:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \text{ approximately } \sim N\Big(0,\ (\hat{I}_1\hat{I}_2^{-1}\hat{I}_1)^{-1}\Big).$$

2.12 Laubach and Williams (2003, REStat)

State-space model and Kalman filter MLE

MLE of the unobservable variables, such as the natural rate of interest and the trend growth rate of output.

Transition Equations:

$$
\begin{aligned}
y_t^* &= y_{t-1}^* + g_t + \varepsilon_{4,t}, \\
g_t &= g_{t-1} + \varepsilon_{5,t}, \\
z_t &= z_{t-1} + \varepsilon_{3,t}, \\
r_t^* &= c g_t + z_t.
\end{aligned}
$$

Measurement Equations:

$$
\begin{aligned}
\tilde{y}_t &= a_{y,1}\tilde{y}_{t-1} + a_{y,2}\tilde{y}_{t-2} + \frac{a_r}{2}\sum_{j=1}^{2}\left(r_{t-j} - r_{t-j}^*\right) + \varepsilon_{t,1}, \\
\pi_t &= B(L)\pi_{t-1} + b_y\tilde{y}_{t-1} + b_i(\pi_t^I - \pi_t) + b_o(\pi_{t-1}^o - \pi_{t-1}) + \varepsilon_{2,t}.
\end{aligned}
$$

MLE using the Kalman filter.

Output

We can compute the MLE of the unobservable variables, such as y^*, g, z, and r^*.

3 Newton Methods for Numerical Optimization

3.1 Newton-Raphson root-finding algorithm

Want. find x s.t. $f(x) = 0$.

Given x_n,

- linearly approximate f at x_n: $f_n = f(x_n) + f'(x_n)(x - x_n)$
- find x_{n+1} s.t. $f_n(x_{n+1}) = 0$
- $x_{n+1} = x_n - [f'(x_n)]^{-1} f(x_n)$.

Figure source: Wikipedia

3.2 Newton's algorithm in optimization

Optimization via Newton–Raphson

Want. find $\arg\min f(x) \iff$ find x s.t. $f'(x) = 0$.

Apply the N-R algorithm to $f'(x) = 0$. Given x_n,

- linearly approximate f' at x_n: $g_n = f'(x_n) + f''(x_n)(x - x_n)$
- find x_{n+1} s.t. $g_n(x_{n+1}) = 0$
- $x_{n+1} = x_n - [f''(x_n)]^{-1} f'(x_n)$.

Alternative interpretation. Given x_n,

- quadratic approximation of f at x_n:

$$g_n = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2} f''(x_n)(x - x_n)^2$$

- find x_{n+1} s.t. $x_{n+1} = \arg\min g_n$.
- f.o.c.:

$$0 = g_n'(x_{n+1}) = f'(x_n) + f''(x_n)(x_{n+1} - x_n).$$

- $x_{n+1} = x_n - [f''(x_n)]^{-1} f'(x_n)$.

Scalar and multivariate cases; computational issues

Scalar case:

$$x_{n+1} = x_n - [f''(x_n)]^{-1} f'(x_n).$$

Multivariate case. $f : \mathbb{R}^k \to \mathbb{R}$.

$$x_{n+1} = x_n - \left[D^2 f(x_n)\right]^{-1} Df(x_n),$$

where $D^2 f(x_n)$ is the Hessian $(k \times k)$ and $Df(x_n)$ is the gradient $(k \times 1)$. Computational issues.

- Computing $H_n = D^2 f(x_n)$ and inverting it is numerically costly.
- Note that at $x^* = \arg\min f$, $H = D^2 f(x^*)$ is p.s.d.
- However, there is no guarantee that H_n is also p.s.d. Thus, we are not sure whether

$$x_{n+1} = \arg\min g_n(x),$$

where

$$g_n(x) = f(x_n) + Df(x_n)(x - x_n) + \frac{1}{2}(x - x_n)' H_n (x - x_n),$$

or whether x_{n+1} is a better approximation of x^* than x_n.

3.3 Quasi-Newton method

Secant idea (scalar)

Want. find x s.t. $f'(x) = 0$.

- "Full" Newton:

$$y = f'(x_n) + f''(x_n)(x - x_n) \Rightarrow x_{n+1} = x_n - [f''(x_n)]^{-1} f'(x_n).$$

- Quasi-Newton (secant method):

$$y = f'(x_n) + \frac{f'(x_n) - f'(x_{n-1})}{x_n - x_{n-1}}(x - x_n) \Rightarrow x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f'(x_n) - f'(x_{n-1})} f'(x_n).$$

Approximate Hessian and remarks

Scalar case:

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f'(x_n) - f'(x_{n-1})} f'(x_n).$$

- Note that

$$f''(x_n) \approx \frac{f'(x_n) - f'(x_{n-1})}{x_n - x_{n-1}} \equiv B_n.$$

- We are approximating $[f''(x_n)]^{-1}$ by using x_n, x_{n-1}, $f'(x_n)$, and $f'(x_{n-1})$.
- w/o directly computing $f''(x_n)$ and w/o inverting it.
- Remarks.
  - Multivariate cases are more complicated.
  - Fast (no 2nd order numerical diff. and matrix inversion).
  - The approximated H_n^{-1} can be restricted to be p.d. (e.g., BFGS algorithm). Thus, $g_n(x_{n+1}) < g_n(x_n)$.
  - The approximated hessian may not be very accurate. That is, $\lim_n B_n$ may not be close to the true hessian $H = D^2 f(x^*)$ even if $x_n \to x^*$.
  - MATLAB: If you need an estimate of the true hessian, use `fminunc`.
  - `fminunc` returns $D^2 f(\lim_n x_n)$, numerical hessian around x^*.
  - `fmincon` returns $\lim_n B_n$.