Problems

1. Discuss the basic differences between the maximum a posteriori and maximum- likelihood estimates of the parameter vector in a linear regression model.
2. Starting with the cost function of $e(w) = e_0(w) + \frac{\lambda}{2}||w||^2$, where,

$$e_0 = \frac{1}{2}\sum_{i=1}^{N}(d_i - w^T x_i)^2,$$

derive the formula of $\widehat{w}_{MAP}(N) = [R_{xx}(N) + \lambda I]^{-1} r_{dx}(N)$ by minimizing the cost function with respect to the unknown parameter vector **w**.

Solution: To derive the **Maximum A Posteriori (MAP) estimate** of the weight vector $\widehat{w}_{\text{MAP}}$, we begin with the **regularized least squares** cost function:

$$e(w) = e_0(w) + \frac{\lambda}{2} \| w||^2$$

where:

- $e_0 = \frac{1}{2}\sum_{i=1}^{N}(d_i - w^T x_i)^2$, is the empirical squared error (data fidelity term),
- $\frac{\lambda}{2} \| w||^2$ is the regularization term (Tikhonov regularization or ridge penalty),
- $\lambda > 0$ is the regularization parameter.

a. Matrix Form of the Cost Function

Let us express $e(w)$ in matrix notation. Define:

- $\boldsymbol{X} \in \mathbb{R}^{N \times M}$: each row is x_i^T,
- $\boldsymbol{w} \in \mathbb{R}^{M \times 1}$: weight vector,
- $\boldsymbol{d} \in \mathbb{R}^{N \times 1}$: desired response vector.

Then:

$$e_0(w) = \frac{1}{2} \| \boldsymbol{d} - \boldsymbol{X}w||^2 \ \ and \ \frac{\lambda}{2} \| w||^2 = \frac{\lambda}{2} w^T w$$

So the total cost becomes:

$$e(w) = \frac{1}{2}(\boldsymbol{d} - \boldsymbol{X}w)^T(\boldsymbol{d} - \boldsymbol{X}w) + \frac{\lambda}{2} w^T w$$

b. Gradient of the Cost Function

To find the optimal $\widehat{w}_{\text{MAP}}$, take the gradient of e(w) with respect to w and set it to zero:

$$\nabla_w e(w) = -\boldsymbol{X}^T(\boldsymbol{d} - \boldsymbol{X}w) + \lambda w$$

Set $\nabla_w e(w) = 0$:

$$\boldsymbol{X}^T\boldsymbol{X}w + \lambda w = \boldsymbol{X}^T\boldsymbol{d}$$

Factor out w:

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda I)w = \boldsymbol{X}^T\boldsymbol{d}$$

c. Solution for w

$$\widehat{w}_{MAP} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{d}$$

d. Interpretation with Autocorrelation Matrices

Let us define:

- $R_{xx}(N) = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$ - input autocorrelation matrix,
- $r_{dx}(N) = \frac{1}{N}\sum_{i=1}^{N} d_i x_i = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{d}$ – cross-correlation between input and desired output.

Then we multiply both numerator and denominator by $\frac{1}{N}$, and we obtain:

$$\widehat{w}_{MAP}(N) = [\boldsymbol{R}_{xx}(N) + \lambda I]^{-1} r_{dx}(N)$$

The **MAP estimate** $\widehat{w}_{MAP}$ is the solution that minimizes the **regularized empirical risk**, balancing **fidelity to data** and **penalization of large weights** (which could overfit). This framework naturally arises in Bayesian linear regression, where the regularization term corresponds to a **Gaussian prior** on **w**. This derivation is fundamental in statistical learning theory and underpins the ridge regression solution.

3. In this problem, we address properties of the least-squares estimator based on the linear regression model of given figure;

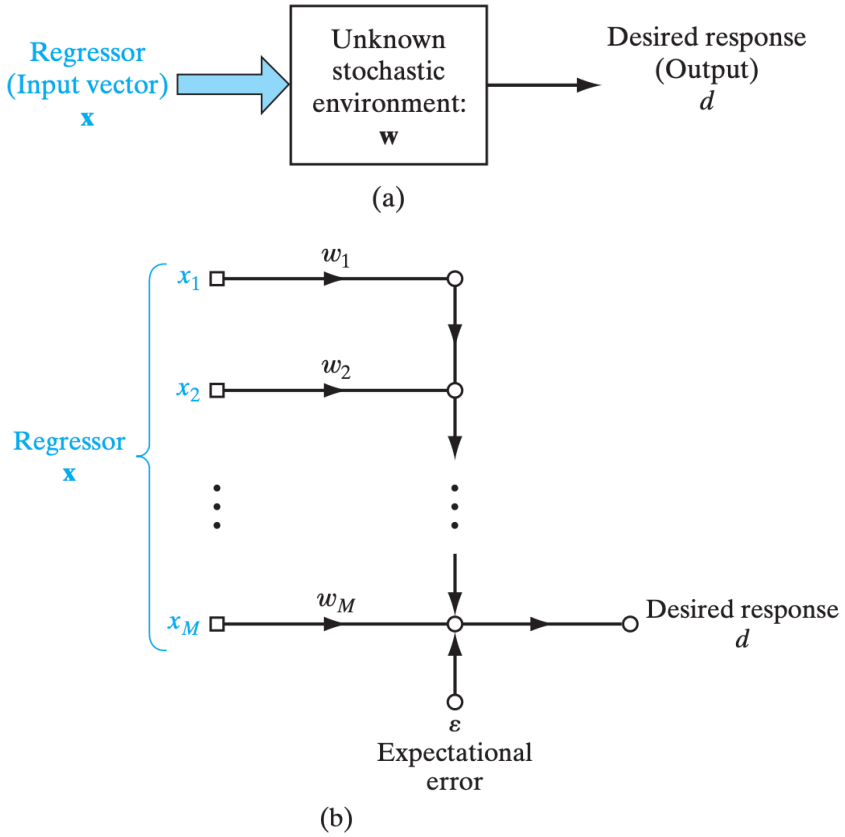

Fig: (a) Unknown stationary stochastic environment. (b) Linear regression model of the environment.

Property 1. The least-squares estimate is unbiased, provided that the expectational error ε in the linear regression model of given figure has zero mean.

$$\hat{w} = \boldsymbol{R_{xx}^{-1} r_{dx}}$$

Property 2. When the expectational error ε is drawn from a zero-mean white- noise process with variance σ^2, the covariance matrix of the least-squares estimate $\hat{w}$ equals

$$\sigma^2 \hat{R}^{-1} xx$$

Property 3. The estimation error

$$e_0 = d - \hat{w}^T \boldsymbol{x}$$

produced by the optimized method of least squares is orthogonal to the estimate of the desired response, denoted by $\hat{d}$; this property is a corollary to the *principle of orthogonality.* If we were to use geometric representations of d, $\hat{d}$, and e_O, then we would find that the

"vector" representing e_O, is perpendicular (i.e., normal) to that representing $\hat{d}$; indeed it is in light of this geometric representation that the formula

$$\hat{R}_{xx}\hat{w} = \hat{r}_{dx}$$

is called the normal equation.

Starting with the normal equation, prove each of these three properties under the premise that $\hat{R}_{\mathrm{XX}}$ and $\hat{r}_{\mathrm{dx}}$ are time-averaged correlation functions.

Solution: Using normal equation,

$$\hat{R}_{xx}\hat{w} = \hat{r}_{dx}$$

$$\hat{w} = \boldsymbol{\hat{R}_{xx}^{-1}\hat{r}_{dx}}$$

Where,
$\hat{R}_{xx} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$ - time-averaged input autocorrelation matrix

$\hat{r}_{dx} = \frac{1}{N}\sum_{i=1}^{N} d_i x_i = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{d}$ - time-averaged cross-correlation between input and output
$d_i = w^T x_i + \varepsilon_i$ - ε_i is zero-mean noise

Property 1: **Unbiasedness of Least Squares Estimate**

Claim: If $E[\varepsilon] = 0$, then $E[\hat{w}] = w$, i.e., $\hat{w}$ is an unbiased estimator of w.

Proof:

From the model: $d_i = w^T x_i + \varepsilon_i \Rightarrow d = \boldsymbol{X}w + \varepsilon$
So,

$$\hat{r}_{dx} = \frac{1}{N}\sum_{i=1}^{N} d_i x_i = \frac{1}{N}\sum_{i=1}^{N}(w^T x_i + \varepsilon_i)x_i = \hat{R}_{xx}w + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i$$

Now take expectation:

$$E[\hat{r}_{dx}] = \hat{R}_{xx}w + E[\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i]$$

Assuming $\varepsilon_i$and x_i are independent and $E[\varepsilon_i] = 0$, the second term is zero.

Thus: $E[\hat{r}_{dx}] = \hat{R}_{xx}w$

Then: $E[\hat{w}] = \hat{R}_{xx}^{-1}E[\hat{r}_{dx}] = \hat{R}_{xx}^{-1}\hat{R}_{xx}w = w$

Proved: The estimator is **unbiased** under zero-mean error.

Property 2: Covariance of Least Squares Estimate

Claim:
If $\varepsilon \sim N(0, \sigma^2 I)$, then:

$$Cov[\hat{w}] = \sigma^2 \hat{R}_{xx}^{-1}$$

Proof:

Recall:

$$\hat{w} = \hat{R}_{xx}^{-1}\hat{r}_{dx} = \hat{R}_{xx}^{-1}(\hat{R}_{xx}w + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i) = w + \hat{R}_{xx}^{-1}(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i)$$

Defining:

$$\delta w = \hat{w} - w = \hat{R}_{xx}^{-1}(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i)$$

Now compute:

$$Cov[\hat{w}] = E[\delta w \delta w^T] = \hat{R}_{xx}^{-1} E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i x_i\right)\left(\frac{1}{N}\sum_{j=1}^{N}\varepsilon_j x_j\right)^T\right]\hat{R}_{xx}^{-1}$$

Because $E[\varepsilon_i \varepsilon_j] = \sigma^2 \delta_{ij}$, this becomes:

$$= \hat{R}_{xx}^{-1}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 x_i x_i^T\right)\hat{R}_{xx}^{-1}$$

$$= \hat{R}_{xx}^{-1}\left(\frac{\sigma^2}{N}\frac{1}{N}\sum_{i=1}^{N}x_i x_i^T\right)\hat{R}_{xx}^{-1}$$

$$= \frac{\sigma^2}{N}\hat{R}_{xx}^{-1}\hat{R}_{xx}\hat{R}_{xx}^{-1}$$

$$= \sigma^2 \hat{R}_{xx}^{-1}$$

Proved: Covariance of the estimator is proportional to $\hat{R}_{xx}^{-1}$.

Property 3: Orthogonality of Estimation Error

Claim:
The error vector $e_0 = d - \hat{d} = d - \boldsymbol{X}\hat{w}$ is orthogonal to the estimated response $\hat{d} = \boldsymbol{X}\hat{w}$. i.e.,

$$e_0{}^T \hat{d} = 0$$

Proof:

Let:

- $\hat{d} = \boldsymbol{X}\hat{w}$
- $e_0 = d - \boldsymbol{X}\hat{w}$

Now compute:

$$e_0{}^T \hat{d} = (d - \boldsymbol{X}\widehat{w})^T \boldsymbol{X}\widehat{w} = d^T \boldsymbol{X}\widehat{w} - \widehat{w}^T \boldsymbol{X}^T \boldsymbol{X}\widehat{w}$$

From the **normal equation**:

$$\boldsymbol{X}^T d = \boldsymbol{X}^T \boldsymbol{X}\widehat{w} \Rightarrow d^T \boldsymbol{X}\widehat{w} = \widehat{w}^T \boldsymbol{X}^T \boldsymbol{X}\widehat{w}$$

So:

$$e_0{}^T \hat{d} = \widehat{w}^T \boldsymbol{X}^T \boldsymbol{X}\widehat{w} - \widehat{w}^T \boldsymbol{X}^T \boldsymbol{X}\widehat{w} = 0$$

Proved: The estimation error is orthogonal to the estimated output vector.

Final Summary

Property	Statement	Result
1. Unbiasedness	$E[\widehat{w}] = w \; if \; E[\varepsilon] = 0$	Proven
2. Covariance	$Cov[\widehat{w}] = \sigma^2 \hat{R}_{xx}^{-1}$ for white noise	Proven
3. Orthogonality	$(d - \boldsymbol{X}\widehat{w})^T \boldsymbol{X}\widehat{w} = 0$	Proven

4. Let $\mathbf{R}_{XX}$ denote the ensemble-averaged correlation function of the regressor **x**, and let $\mathbf{r}_{dx}$ denote the corresponding ensemble-averaged cross-correlation vector between the regressor **x** and response *d*; that is,

$$\boldsymbol{R}_{xx} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$$
$$\boldsymbol{r}_{dx} = \mathbb{E}[d\boldsymbol{x}]$$

Referring to the linear regression model of Eq. $d = w^T \boldsymbol{x} + \epsilon$, show that minimization of the mean- square error leads to the *Wiener–Hopf equation*

$$\boldsymbol{J}(\boldsymbol{w}) = \mathbb{E}[\epsilon^2]$$

$$\boldsymbol{R}_{xx}\boldsymbol{w} = \boldsymbol{r}_{dx}$$

where **w** is the parameter vector of the regression model. Compare this equation with the normal equation of Eq. $\hat{R}_{xx(N)}\widehat{w}(N) = \hat{r}_{dx(N)}$.

Solution: The **linear regression model**:

$$d = w^T x + \varepsilon$$

Where:

- $\boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{M}}$: regressor (random vector)
- $\boldsymbol{w} \in \boldsymbol{R}^{\boldsymbol{M}}$: parameter vector

- d: desired scalar response
- ε: error (zero-mean random variable)

We want to minimize the **mean square error (MSE)**:

$$\boldsymbol{J(w) = E[\varepsilon^2] = E[(d - w^T x)^2]}$$

Expanding the cost function:

$$J(w) = E[(d - w^T x)^2] = E[d^2 - 2dw^T x + w^T x x^T w]$$

Using linearity of expectation:

$$J(w) = E[d^2] - 2w^T E[dx] + w^T E[xx^T] w$$

Define:

- $r_{dx} = E[dx]$ — cross-correlation vector
- $R_{xx} = E[xx^T]$ — autocorrelation matrix

Then:

$$J(w) = E[d^2] - 2w^T r_{dx} + w^T R_{xx} w$$

Minimize $J(w)$

To find the optimal w, take the gradient of $J(w)$ and set it to zero:

$$\nabla_w J(w) = -2r_{dx} + 2R_{xx} w = 0$$

Solve for w:

$$R_{xx} w = r_{dx}$$

$\Rightarrow$ This is the **Wiener–Hopf equation**.

Connection with MSE

Recall that:

$$J(w) = E[(d - w^T x)^2] = E[\varepsilon^2]$$

So once w is optimized by solving the Wiener–Hopf equation, the cost $J(w)$ is minimized and equals the **minimum mean square error**.

Comparison with the Normal Equation

- **Wiener–Hopf equation** (theoretical/ensemble):

$$R_{xx}w = r_{dx}$$

- **Normal equation** (empirical/time-averaged):

$$\hat{R}_{xx}(N)\hat{w}(N) = \hat{r}_{dx}(N)$$

Comparison:

Aspect	Wiener–Hopf Equation	Normal Equation
Type	Theoretical / Expectation	Empirical / Sample-Based
Uses	Ensemble averages	Time averages
Equation	$R_{xx}w = r_{dx}$	$\hat{R}_{xx}(N)\hat{w}(N) = \hat{r}_{dx}(N)$
Application	Ideal if distribution is known	Practical estimation from data

Conclusion: Thus, the **normal equation** is the finite-sample approximation to the **Wiener–Hopf equation**, which holds in the infinite-sample (ensemble) case.

5. Equation

$$L_{\text{av}}(f(\mathbf{x}), F(\mathbf{x}, \mathcal{T})) = B^2(\hat{\mathbf{w}}) + V(\hat{\mathbf{w}})$$

expresses the natural measure of the effectiveness of the approximating function $F(\boldsymbol{x}, \hat{\boldsymbol{w}})$ as a predictor of the desired response d. This expression is made up of two components, one defining the squared bias and the other defining the variance. Derive this expression, starting from Eq.

$$L_{\text{av}}(f(\mathbf{x}, \mathbf{w}), F(\mathbf{x}, \hat{\mathbf{w}})) = \mathbb{E}_{\mathcal{T}}[(\mathbb{E}[d|\mathbf{x}] - F(\mathbf{x}, \mathcal{T}))^2]$$

Solution: The **expected squared error** is

$$E_{D,x}[\left(d - \hat{F}(x)\right)^2]$$

Where:

- $\widehat{\boldsymbol{F}}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}, \widehat{\boldsymbol{w}})$: is the learned function from data D (e.g., least squares solution).
- d: desired output (a random variable, often modeled as $\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}, with\ \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2})$
- The expectation is over both the data D used to train the model and over new inputs x.

Step 1: Start with the expected prediction error

Let the true model be:

$$d = f(x) + \varepsilon, with\ E[\varepsilon] = 0, Var[\varepsilon] = \sigma^2$$

Now consider the expected squared prediction error:

$$E_{D,x,\varepsilon}[\left(f(x)+\varepsilon-\hat{F}(x)\right)^2]$$

This can be expanded as:

$$E_x[E_{D,\varepsilon}[\left(f(x)+\varepsilon-\hat{F}(x)\right)^2 \mid x]]$$

Step 2: Decompose the error

Let's fix x and examine:

$$E_{D,\varepsilon}[\left(f(x)+\varepsilon-\hat{F}(x)\right)^2$$

This becomes:

$$= E_{D,\varepsilon}[\left(\left(f(x)-E\left[\hat{F}(x)\right]\right)+\left(E\left[\hat{F}(x)\right]-\hat{F}(x)\right)+\varepsilon\right)^2]$$

Now let:

- $\text{Bias}(x) = E[\hat{F}(x)] - f(x)$
- $\text{Var}(\hat{F}(x)) = E_D[\left(\hat{F}(x) - E\left[\hat{F}(x)\right]\right)^2]$

Using the identity:

$$E[(a+b+c)^2] = E[a^2]+E[b^2]+E[c^2]+2E[ab]+2E[ac]+2E[bc]$$

Here:

- $a = f(x) - E[\hat{F}(x)] = -Bias(x)$ (a constant w.r.t. D)
- $b = E[\hat{F}(x)] - \hat{F}(x)$ (zero-mean random variable)
- $c = \varepsilon$ (independent noise)

Cross terms vanish due to independence and zero mean:

$$\Rightarrow E[\left(d-\hat{F}(x)\right)^2] = Bias(x)^2 + Var(\hat{F}(x)) + \sigma^2$$

Final Expression: Bias-Variance Decomposition

$$E[\left(d-\hat{F}(x)\right)^2] = \left(Bias\left[\hat{F}(x)\right]\right)^2 + Var[\hat{F}(x)] + \sigma^2$$

Taking the expectation over x, we get the full average prediction error:

$$E_x[E_{D,\varepsilon}[\left(d - \hat{F}(x)\right)^2]] = E_x[Bias(x)^2 + Var(\hat{F}(x))] + \sigma^2$$

Interpretation:

Term	Meaning
$\boldsymbol{Bias^2}$	Error due to approximating the true function with a simplified model (underfitting)
$\boldsymbol{Variance}$	Error due to model sensitivity to training data (overfitting)
$\boldsymbol{\sigma^2}$	Irreducible noise (inherent randomness in the data)

This decomposition shows the **tradeoff** between **bias** and **variance** in machine learning models.

6. Elaborate on the following statement:

A network architecture, constrained through the incorporation of prior knowledge, addresses the bias–variance dilemma by reducing variance at the expense of increased bias.

Solution: The Bias–Variance Dilemma

The **bias–variance trade-off** is central to understanding generalization in machine learning. For a given input x, the expected prediction error of a model can be decomposed as:

$$E\left[\left(d - \hat{F}(x)\right)^2\right] = \underbrace{(Bias[x])^2}_{underfitting} + \underbrace{Variance[x]}_{overfitting} + \underbrace{\sigma^2}_{irreducible\ noise}$$

- **High bias** means the model is too simple to capture the data complexity → **underfitting**.
- **High variance** means the model is too flexible and overfits training data noise → **overfitting**.

The goal is to **balance bias and variance** to minimize prediction error on unseen data.

What Does *"Incorporation of Prior Knowledge"* Mean?

Incorporating **prior knowledge** refers to **embedding domain-specific constraints or assumptions** into the model architecture or learning process.

Examples:

- *Using **convolutional layers** in CNNs for image tasks (exploiting spatial locality and translation invariance).*
- *Using **recurrent structures** for time-series data.*
- *Designing **sparse connectivity** in the network.*
- *Applying **weight sharing**, **symmetries**, or **invariance constraints**.*
- *Preferring **simpler architectures** or **regularization**.*

These constraints reflect assumptions like:

- *Only local pixels matter (CNNs),*
- *The sequence matters (RNNs),*
- *Simpler models are better (Occam's razor).*

How Prior Knowledge Affects Bias and Variance

⇒ **Reduces Variance:** By **limiting the hypothesis space** (i.e., the set of functions the model can represent), the model becomes **less sensitive to fluctuations in the training data**. That is:

- *Less likely to overfit,*
- *More stable predictions across different datasets.*

This improves **generalization** and reduces **variance**.

⇒ **Increases Bias:** However, if the **assumptions are too strong or incorrect**, the model **may not be able to fit the data well**, even with unlimited training data. That is:

- *The model is **constrained to learn only a subset of possible patterns**,*
- *It may **miss true relationships** if they fall outside the prior constraints.*

This increases **bias**.

Summary of the Trade-off

Constraint	Bias	Variance	Effect
No constraints (high-capacity model)	Low	High	Overfits noisy data (high variance)
With strong prior knowledge (constrained model)	High	Low	Underfits if assumptions are too strong

By **intelligently incorporating prior knowledge**, we **reduce variance** and achieve better generalization — **as long as the increase in bias is not too costly**.

Final Interpretation

So, the original statement means:

Constraining a model (e.g., by architecture or regularization) to follow certain known behaviors or patterns helps **prevent overfitting** by reducing how much the model's predictions change in response to different training datasets (i.e., reducing variance). However, this comes at the cost of potentially **oversimplifying the model**, which may prevent it from capturing all nuances of the data (i.e., increasing bias).
This is the essence of regularization, model design, and informed architecture choices.

7. The method of instrumental variables described in Eq.

$$\hat{\mathbf{w}}(N) = \mathbf{R}_{z\hat{x}}^{-1}\mathbf{r}_{d\hat{x}} = \left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i\mathbf{z}_i^T\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i d_i\right)$$

provides an asymptotically unbiased estimate of the unknown parameter vector $\widehat{\boldsymbol{w}}$(N); that is,

$$\lim_{N\to\infty}\widehat{\boldsymbol{w}}(N) = \boldsymbol{w}$$

Prove the validity of this statement, assuming joint ergodicity of the regressor **x** and response *d*.

Solution: To **prove that the instrumental variables (IV) method provides an asymptotically unbiased estimate** of the parameter vector $\widehat{w}(N)$, i.e.,

$$\lim_{N\to\infty}\widehat{w}\,(N) = w$$

we start by understanding the **IV method**, and then prove this convergence **under the assumption of joint ergodicity** of the regressor x and response d.

1. Background: Linear Regression Model

We consider the **linear model**:

$$d(n) = w^T x(n) + \varepsilon(n)$$

Where:

- $x(n) \in R^M$: regressor (possibly correlated with noise)
- $d(n) \in R$: scalar desired output
- $w \in R^M$: unknown parameter vector
- $\varepsilon(n)$: zero-mean noise (possibly correlated with x(n))

Problem with Ordinary Least Squares (OLS)

The OLS estimator:

$$\widehat{w}OLS(N) = = \left(\frac{1}{N}\sum_{n=1}^{N}x(n)x(n)^T\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}x(n)d(n)\right)$$

can be **biased** if $\varepsilon(n)$ is correlated with x(n). That is, if:

$$E[x(n)\varepsilon(n)] \neq 0$$

Instrumental Variables (IV) Estimator

To fix this, the IV method uses an **instrumental variable** z(n) that satisfies:

1. $E[z(n)\varepsilon(n)] = 0$ (uncorrelated with noise)
2. $E[z(n)x(n)^T]$ is full rank

Then the **IV estimator** is:

$$\widehat{w}_{IV}(N) = = \left(\frac{1}{N}\sum_{n=1}^{N} z(n)x(n)^T\right)^{-1} (\frac{1}{N}\sum_{n=1}^{N} x(n)d(n))$$

This is often written as:

$$\widehat{w}_{IV}(N) = R_{zx}^{-1}(N)\hat{r}_{zd}(N)$$

Where:

- $\hat{R}_{zx}(N) = \frac{1}{N}\sum_{n=1}^{N} z(n)x(n)^T$
- $\hat{r}_{zd}(N) = \frac{1}{N}\sum_{n=1}^{N} z(n)d(n)$

Assumption: Joint Ergodicity

Joint ergodicity of $\{x(n), d(n), z(n)\}$ implies:

$$\lim_{N\to\infty} \hat{R}_{zx}\,(N) = E[z(n)x(n)^T] = R_{zx}$$
$$\lim_{N\to\infty} \hat{r}_{zd}\,(N) = E[z(n)d(n)] = r_{zd}$$

So in the limit:

$$\lim_{N\to\infty} \widehat{w}_{IV}(N) = R_{zx}^{-1}(N)r_{zd}(N)$$

Prove Asymptotic Unbiasedness

Recall that $d(n) = w^T x(n) + \varepsilon(n)$, so:

$$r_{zd} = E[z(n)d(n)] = E\big[z(n)\big(w^T x(n) + \varepsilon(n)\big)\big] = E[z(n)x(n)^T]w + E[z(n)\varepsilon(n)]$$

Now, by assumption of **instrument validity**:

$$E[z(n)\varepsilon(n)] = 0$$

So:

$$r_{zd} = R_{zx}w$$

Then:

$$\lim_{N\to\infty} \widehat{w}_{IV}(N) = R_{zx}^{-1}(N)r_{zd} = R_{zx}^{-1}(N)(N)R_{zx}w = w$$

Proved: The IV estimator is **asymptotically unbiased**, i.e.,

$$\lim_{N\to\infty} \widehat{w}_{IV}(N) = w$$

under:

- Valid instruments (uncorrelated with noise),
- Full-rank R_{zx},
- Joint ergodicity of $x(n), d(n), z(n)$.

Summary Table

Condition	Role
$\boldsymbol{E[z(n)\varepsilon(n)] = 0}$	Instruments are uncorrelated with noise
$\boldsymbol{E[z(n)x(n)T]}$ is full-rank	Invertibility of R_{zx}
Joint ergodicity	Ensures convergence of time averages to expectations