Math Review Notes—Statistical Learning

Gregory Faletto

Contents

Last updated July 31, 2020

Chapter 1

Statistical Learning

These notes are based on my notes from Math 547: Mathematics of Statistical Learning at USC taught by Steven Heilman, GSBA 604: Regression and Generalized Linear Models for Business Applications at USC taught by Gourab Mukherjee, and DSO 677: Dynamic Programming and Markov Decision Processes taught by Paat Rusmevichientong, which used the textbooks [Bertsekas, 2012a] and [Bertsekas, 2012b]. I also borrowed from some other sources which I mention when I use them.

1.1 Segmented regression, local regression, splines

Broken stick regression/segmented regression: useful if data has two different groups.

$$y = B_\ell(x) + B_r(x) = \beta_0 + \beta_1(c-x)_+ + \beta_2(x-c)_+$$

where c is the breaking point and is fixed (if you use data to choose c, becomes a nonlinear problem). Advantages: more localized

Splines: like combination of broken stick regression and polynomial regression. fit a polynomial in each segment.

for splines: knot selection is important (bias/variance tradeoff)

1.1.1 Local Regression

Local regression: weight chosen by kernel:

$$w_i = \frac{K((x_i - x_0)/\sqrt{v})}{\sum_{i=1}^{M(s)} K((x_i - x_0)/\sqrt{v})}$$

$$\text{dnorm}(x_i, \text{mean} = x_0; \text{sd} = \sqrt{v})$$

$$K((x_i - x_0)/\sqrt{v}) = \phi((x_i - x_0)/\sqrt{v}) \cdot v^{-1/2} = \frac{1}{\sqrt{2\pi v}} \exp - \left(\frac{(x_i - x_0)^2}{2v} \right)$$

so if the distance between two points increases than the weight decreases. This is called kernel density with a Gaussian kernel.

$$x = x_0; \hat{\beta}(x_0); \hat{\sigma}(x_0)$$

Two tuning parameters: v (bandwidth) and s. Disadvantages: hard to interpret, starts to do poorly in high dimensions.

1.1.2 Curse of Dimensionality (brief discussion)

Suppose $x_i : i \in [n] \sim F$ i.i.d. with fixed dimension, and F has bounded support $[a, b]^d$. Let $x \in \text{supp}(F)$. Then

$$\min_x d_2(x, x_i) \sim \frac{1}{n^{1/d}}$$

so goes to 0 as $n \to \infty$ for fixed d. But bad in d; for example, if $d = 5$, need $n = 10^5$ for $d = 0.1$. As this minimum distance increases, regularity becomes more difficult.

1.2 Dimension Reduction methods

1.2.1 Principal components regression

(See Section **??** for more details on principal components.) Given data (y, X), don't look at y for now. Look for maximum variability direction of X:

$$P_1 = \underset{\|a\|_2=1}{\arg\max} \text{Var}(a^T x) = a^T (X^T X) a.$$

then (letting $\mathcal{P}_1$ be the projection matrix for projecting onto P_1)

$$P_2 = \underset{\|a\|_2=1, \mathcal{P}_1 a=0}{\arg\max} \text{Var}(a^T x) = a^T (X^T X) a.$$

and so on. We will regress y against the first M principal components of X for some $X \leq p$, where the principal components are the columns of U:

$$\hat{y}^{\text{pcr}}_{(M)} = \overline{y} + \sum_{m=1}^{M} \hat{\theta}_m z_m,$$

where $\hat{\theta}_m = (z_m^T y)/(z_m^T z_m)$ since all the z_m are orthogonal. Also, since the z_m are linear combinations of the original x_j, the solution can be expressed in terms of coefficients in the original feature space:

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m,$$

Works well when high variability directions in the explanatory variables are also interesting attributes to study.

Sometimes if data are in a mainfold that isn't linearly parsed, good to use an ISOMAP or LLE (local linear embedding).

(read: starting on p.63 of ESL)

1.2.2 Partial least squares

Does look at y, unlike PCR. $X_{n\times p} \to W_{n\times 3}$. Start by standardizing all x_j to have mean 0 and unit variance. Then compute $\hat{\phi}_{1j} = \langle x_j, y\rangle$ for each j. Then the derived input $z_1 = \sum_j \hat{\phi}_{1j} x_j$ is the first partial least squares direction. Then y is regressed on z_1 giving coefficient $\hat{\theta}_1$, and then $x_1, \ldots, x_j$ are orthogonalized with respect to z_1. Continue until $M \leq p$ directions have been obtained. (see p. 80 of ESL)

$$\vdots$$

notes from GSBA 604: First reduced dimension:

$$w_{n\times 1}^{(1)} = X_{n\times 1}^{(1)} + X_{n\times 1}^{(2)} + \ldots + X_{n\times 1}^{(p)}$$

proportional correlation $(y, x^{(1)})$ to

1.2.3 Dimension reduction by random matrix

Let $R \in \mathbb{R}^{p\times d}$, $d < p$, with $R_{ij} \sim \mathcal{N}(0,1)$. Let $\tilde{X} = XR$. By the Johnson-Lindenstrauss lemma, for any $\epsilon > 0$ there exists an R such that

$$(1-\epsilon)\|\tilde{X}_i - \tilde{X}_j\|_2^2 \leq \|X_i - X_j\|_2^2 \leq (1+\epsilon)\|\tilde{X}_i - \tilde{X}_j\|_2^2.$$

(isometric transformation—distances are maintained)

1.3 Goodness of fit, residuals, residual diagnostics, leverage

Goodness of fit: F test. Assume $\text{Var}(\epsilon) = \sigma^2 I$ and recall $\hat{\epsilon} = (I - H)\epsilon$. So $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H)$ and $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - H_{ii})$. So $h_i := H_{ii}$ is called the **leverage** for the ith case. Some properties: if X is **ill-**

conditioned (condition number—ratio of largest to smallest eigenvalue of Gram matrix $X^T X$—is high), then H_{ii} will vary a lot.

Properties: $\sum_i h_i = p + 1$ (because H_{ii} is idempotent, its trace is equal to $p + 1$—all its eigenvalues are 0 or 1, and it is nonsingular, so all of its eigenvalues are 1). And, all $h_i \geq 1/n$.

Larger h_i result in smaller $\text{Var}(\hat{\epsilon}_i)$, which forces the fit to be close to y_i. Average leverage: $(p + 1)/n$. Rule of thumb: leverages of more than $2(p+1)/n$ should be looked at closely (they have a large influence on the slope, so if they are incorrect then it's a big problem for the fit).

Standardized or Studentized results:

$$r_i = \frac{\hat{\epsilon}_i}{\text{se}(\hat{\epsilon}_i)} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Let $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ be the estimates from the regression with the ith case excluded. Let x_i^T denote the ith row of X, write $X_{(i)}$ for X without the ith row.

$$\vdots$$

Outlier test: can test for outliers using the fact that each $t_i \sim t(n - p - 2)$ if no outliers are present. Need to do a multiplicity correction for multiple testing (say Bonferroni correction). have to do residual diagnostics.

1.3.1 Residual diagnostics

Consider leverage points (outliers, influential points: points that change the slope a lot).

1. **Partial residual plot:** kth variable. $\hat{\epsilon}^{(k)} = y - \sum_{j \neq k} \beta_j x_j$: residuals without the kth explanatory variable. Plot this against x_k and regress. If the resulting slope $\hat{\beta}_k$ is large, this is an influential point.
2. **Added variable plot:** Regress x_k by other explanatory variables $\hat{\delta}^{(k)}$. Plot $\hat{\epsilon}^{(k)}$ against $\hat{\delta}^{(k)}$.

1.4 DSO 607

Generalized linear models:

$$f_n(z, \beta) = \prod_{i=1}^{n} \exp\left[\theta_i z_i - b(\theta_i) h(z_i)\right], \quad z = (z_1, \ldots, z_n)^T$$

Natural parameter θ_i: $\theta_i = x_i^T \beta$, $x_i = \{x_{ij} : j \in \mathscr{M}\}$

$h(z_i)$: normalization constant

linear regression: $b(\theta) = \frac{1}{2}\theta^2$

other: $b(\theta) = \log(1 + e^\theta)$

If $Y = (Y_1, \ldots, Y_n)^T \sim F_n(\cdot, \beta)$, then $\mathbb{E}(Y) = (b'(\theta_1), \ldots, b'(\theta_n))^T = \mu(\theta)$ and

$\text{Cov}(Y) = \text{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\} = \Sigma(\theta)$ where $\theta = X\beta$ and $X = (x_1, \ldots, x_n)^T$ is the $n \times d$ design matrix.

Quasi-log-likelihood ("quasi" because error may be misspecified):

$$\ell_n(y, \beta) = y^T X\beta - \mathbf{1}^T b(X\beta) + \mathbf{1}^T h(y)$$

Like MLE, maximizing $\ell_n(y, \beta)$ with respect to β gives the quasi-MLE $\hat{\beta}_n$. Solution exists and is unique due to strict convexity of b, solves the score equation

$$\frac{\partial \ell_n(y, \beta)}{\partial \beta} = x^T[y - \mu(X\beta)] = \mathbf{0}$$

(Intuition of score equation: the columns of X are all orthogonal to the errors (uncorrelated if X is random).

1.4.1 Akaike Information Criterion (AIC)

AIC: proposed by Akaike [1973] to choose a model by minimizing the Kullback-Leibler (KL) divergence of the fitted model from the true model (or equivalently, maximize the expected log-likelihood). Recall the KL Divergence

$$I(\theta; \theta_0) := 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right)\right] - 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta)\right)\right].$$

We will try to maximizing the KL Divergence by estimating θ_0 as best as we can by maximizing the **probabilistic negentropy**

$$\mathbb{E}_Z I(\theta; \hat{\theta}_0(Z)) := 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right] - 2\mathbb{E}_{\theta_0, Z}\left[\log\left(f\left(X \mid \hat{\theta}_0(Z)\right)\right)\right].$$

Because the true model θ_0 is unknown we cannot carry out this maximization directly. Note that as the number of independent observations increases, the **mean log-likelihood ratio**

$$\hat{I}(\theta; \theta_0) := \frac{2}{n}\sum_{i=1}^{n} \log \frac{f(x_i \mid \theta_0)}{f(x_i \mid \theta)} \xrightarrow{p} I(\theta; \theta_0).$$

Because of this, Akaike reasons that maximizing the mean log-likelihood ratio over θ_0 (i.e. computing the maximum likelihood estimate) tend to maximize the entropy. So the maximum likelihood estimate $\hat{\theta}_0(Z)$ is substituted for the unknown θ_0.

Way we wrote KL Divergence in DSO 607: density f from density g:

$$I(g_n; f_n(\cdot, \beta)) = \int [\log g(z)] g(z) dz - \int [\log f(z)] g(z) dz$$

Akaike [1973] found that up to an additive constant, the KL divergence of the fitted model from the true model can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \dim(\hat{\theta}) = -\ell_n(\hat{\theta}) + \lambda \sum_{j=1}^{p} \mathbf{1}_{\{\hat{\theta}_j \neq 0\}}$$

where $\ell_n(\theta)$ is the log-likelihood function and $\lambda = 1$. This leads to the Akaike information criterion (AIC) for comparing models:

$$AIC\left(\hat{\theta}_k(Z)\right) := n\hat{I}\left(\hat{\theta}_k(Z; \hat{\theta}_0(Z)\right) + 2\|\hat{\theta}_k(Z)\|_0 = 2\sum_{i=1}^{n} \log \frac{f(x_i \mid \hat{\theta}_0(Z))}{f(x_i \mid \hat{\theta}_k(Z))} + 2\|\hat{\theta}_k(Z)\|_0$$

Way we wrote this is DSO 607:

$$AIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + 2\|\hat{\theta}\|_0$$

Intuition: $\log g(x)$ is the log likelihood. Penalty term can be interpreted as penalty, or as a bias correction since you are doing training and feature selection simultaneously on the same data.

$$I(g_n; f_n(\cdot, \beta)) = \sum_{i=1}^{n} \left[\int \right]$$

To minimize the KL divergence

$$\frac{\partial I(g_n; f_n(\cdot, \beta))}{\partial \beta} = -X^T[\mathbb{E}(Y) = \mu(X\beta)] = 0$$

the inverse of the Fisher information matrix is the covariance of the MLE (?).

$$\vdots$$

(For more information on KL Divergence, see Sections **??** and **??**). For AIC, we minimize the KL divergence. For BIC, we maximize the Bayes factor (posterior probability for the model).

1.4.2 Bayesian Information Criterion (BIC)

A typical Bayesian model selection procedure is to first give nonzero prior probability α_M om each model M and then prescribe a prior distribution μ_M for the parameter vector in the corresponding model. The

Bayesian principle of model selection is to choose the most probable model *a posteriori*; that is, to choose a model that maximizes the log-marginal likelihood (or the Bayes factor)

$$\log \int \alpha_M \exp[\ell_n(\theta)] d\mu_m(\theta).$$

Schwarz [1978] took a Bayesian approach with prior distributions that have nonzero prior probabilities on some lower dimensional subspaces of $\mathbb{R}^p$ and showed that the negative log-marginal likelihood can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \|\hat{\theta}\|_0$$

where $\lambda = (\log n)/2$. This asymptotic expansion leads to the Bayesian information criterion (BIC) for comparing models:

$$BIC(\hat{\theta}) := -2 \log \Big(f(x \mid \hat{\theta}; \hat{\theta}_{MLE}) \Big) + (\log n) \|\hat{\theta}\|_0.$$

where f is the density function parameterized by $\hat{\theta}_{MLE}$, the maximum likelihood estimate for the density given the data x.

Way we wrote this in DSO 607:

$$BIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + (\log n) \|\hat{\theta}\|_0.$$

$$\vdots$$

$$B_n^{1/2} A_n (\hat{\beta}_n - \beta_{n,0}) = W_n \xrightarrow{D} \mathcal{N}(0, I_d)$$

$$\hat{\beta}_n - \beta_{n,0} = A_n^{-1} B_n^{1/2} W_n \implies \text{Cov}(\hat{\beta}_n) = \text{Cov}(\hat{\beta} - n - \beta_{n,0})$$

$$= \text{Cov}(A_n^{-1} B^{1/2} W_n) = A_n^{-1} B_n^{1/2} \text{Cov}(W_n) B_n^{1/2} A_n^{-1} = A_n^{-1} B_n^{1/2} I_d B_n^{1/2} A_n^{-1} = \boxed{A_n^{-1} B_n A_n^{-1}}$$

Note that if the model is correct, $A_n = B_n$ so this reduces to conventional asymptotic MLE theory ($\text{Cov}(\hat{\beta}_n) = A_n^{-1}$).

$$\vdots$$

A_n from working model, B_n from true model (unknown).

GBIC in misspecified models: $H_n = A_n^{-1}B_n$ (covariance contrast matrix). Note that when model is specified, $H_n = I_d$ so the log of its determinant is 0 so it vanishes. If not, then it is a misspecification penalty.

$$\vdots$$

Note: $\log(y, \hat{\beta}_n) > \log(y, \beta_{n,0})$ because $\hat{\beta}_n$ is by definition the MLE on the observed data. But $\mathbb{E}(\log(\tilde{y}, \beta_{n,0}) > \mathbb{E}(\log(\tilde{y}, \hat{\beta}_n)$ because $\beta_{n,0}$ is the true parameter. We have a systematic upward bias when we use the empirical estimate. (p.18 of week 2-2 slides)

Proposition 1.4.2.1 (**Result from *"Econometrics: Methods and Applications"* homework**). Consider the usual linear model, where $y = X\beta + \epsilon$. Suppose we compare two regressions, which differ in how many variables are included in the matrix X. In the full (unrestricted) model p_1 regressors are included. In the restricted model only a subset of $p_0 < p_1$ regressors are included. Then for large n, selection based on AIC corresponds to an F-test with a critical value of approximately 2.

Proof. Let e_R be the vector of residuals for the restricted model with p_0 parameters and e_U the vector of residuals for the full unrestricted model with p_1 parameters. Then we have the sample standard deviations

$$s_0^2 = \frac{1}{n-p_0} e_R{}' e_R, s_1^2 = \frac{1}{n-p_1} e_U{}' e_U \tag{1.1}$$

Recall the AIC:

$$\log(s^2) + \frac{2k}{n}$$

where k is the number of regressors included in the model.
For the small model, we have

$$AIC_0 = \log(s_0^2) + \frac{2p_0}{n}.$$

For the big model, we have

$$AIC_1 = \log(s_1^2) + \frac{2p_1}{n}.$$

Therefore the smallest model is preferred according to the AIC if

$$AIC_0 < AIC_1$$

$$\iff \log(s_0^2) + \frac{2p_0}{n} < \log(s_1^2) + \frac{2p_1}{n} \iff \log(s_0^2) - \log(s_1^2) < \frac{2p_1}{n} - \frac{2p_0}{n} \iff \log(\frac{s_0^2}{s_1^2}) < \frac{2}{n}(p_1 - p_0)$$

$$\iff \frac{s_0^2}{s_1^2} < e^{\frac{2}{n}(p_1 - p_0)} \tag{1.2}$$

If n is very large, $\frac{2}{n}(p_1 - p_0)$ is small. Therefore, using the first order Taylor approximation $e^x \approx 1 + x$ we can approximate that

$$e^{\frac{2}{n}(p_1 - p_0)} \approx 1 + \frac{2}{n}(p_1 - p_0)$$

(if n is very large.) Substituting this expression into the right side of (1.2) yields

$$\frac{s_0^2}{s_1^2} < 1 + \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - 1 < \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - \frac{s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$$

$$\iff \frac{s_0^2 - s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$$

for n very large. Plugging in the expressions from (1.1), we have

$$\frac{\frac{1}{n-p_0} e_R{}' e_R - \frac{1}{n-p_1} e_U{}' e_U}{\frac{1}{n-p_1} e_U{}' e_U} < \frac{2}{n}(p_1 - p_0).$$

For large values of n, $n - p_0 \approx n - p_1 \approx n$. This yields

$$\frac{\frac{1}{n} e_R{}' e_R - \frac{1}{n} e_U{}' e_U}{\frac{1}{n} e_U{}' e_U} < \frac{2}{n}(p_1 - p_0)$$

$$= \frac{e_R{}' e_R - e_U{}' e_U}{e_U{}' e_U} < \frac{2}{n}(p_1 - p_0) \tag{1.3}$$

Now recall the F statistic:

$$F = \frac{(e_R{}' e_R - e_U{}' e_U)/g}{e_U{}' e_U/(n-k)} \tag{1.4}$$

where k is the number of explanatory factors in the unrestricted model, and g is the number of explanatory factors removed from the unrestricted model to create the restricted model. Under this test, we believe there is significant evidence to suggest that $\beta \neq 0$ (so the unrestricted model is preferred) if $F > F_{critical}$. Therefore a larger model is preferred if $F > F_{critical}$, and we stay with (prefer) a smaller model if $F < F_{critical}$.

Let $F_{critical} = 2$. Then a smaller model is preferred if $F < 2$:

$$\frac{(e_R{}' e_R - e_U{}' e_U)/g}{e_U{}' e_U/(n-k)} < 2$$

In this case, with p_1 factors in the unrestricted model and p_0 in the restricted model, we get

$$\frac{(e_R{}' e_R - e_U{}' e_U)/(p_1 - p_0)}{e_U{}' e_U/(n-p_1)} < 2$$

$$\frac{(e_R{}' e_R - e_U{}' e_U)}{e_U{}' e_U} < \frac{2(p_1 - p_0)}{n - p_1}$$

If n is very large, $n - p_1 \approx n$. Substituting this in yields

$$\frac{(e_R{}'e_R - e_U{}'e_U)}{e_U{}'e_U} < \frac{2(p_1 - p_0)}{n} \tag{1.5}$$

which equals (1.3). Our condition for preferring a restricted model when doing an F-test with $F_{critical} = 2$ (and when n is very large) is approximately the same as our condition for preferring a restricted model when using the AIC (when n is very large).

□

1.5 Ridge Regression

If p is large, it tends to be better to shrink the last squares estimator. (Even though this introduces bias, it will likely reduce variance, and the tradeoff will often help for some amount of shrinkage.) This is related to the Stein estimator.

Suppose $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, and for all $1 \leq i \leq n$, there are known vectors $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$. Our observed data are $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$. Let $\boldsymbol{X}$ be the $n \times p$ matrix so that the i^{th} row of $\boldsymbol{X}$ is the row vector $x^{(i)}$. Assume that $p \leq n$ and the matrix $\boldsymbol{X}$ has full rank. Let $\lambda > 0$ and consider the quantity

$$\sum_{i=1}^{n} \left(y_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{1.6}$$

The term $\|\boldsymbol{\beta}\|_2^2$ penalizes $\boldsymbol{\beta}$ from having large entries. By Lagrange Multipliers, a critical point $\boldsymbol{\beta}$ of the constrained minimization problem

$$\text{minimize } \sum_{i=1}^{n} (\boldsymbol{y}_i - \langle x^{(i)}, \boldsymbol{\beta} \rangle)^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_2^2 \leq 1$$

is equivalent to the existence of a $\lambda \in \mathbb{R}$ such that $\boldsymbol{\beta}$ is a critical point of (1.6). We call the $\hat{\beta}$ that minimizes (1.6) the **ridge regression** estimator for β.

Proposition 1.5.0.1 (Math 541A Homework Problem). The value of $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that minimizes (1.6) is $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Proof.

$$\begin{aligned}\sum_{i=1}^{n} \left(y_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|^2 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\end{aligned}$$

where $\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ because a scalar equals its transpose. Differentiating with respect to $\boldsymbol{\beta}$ yields

$$\begin{aligned}-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T = 0 &\iff \boldsymbol{\beta}^T(2\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}_p) = 2\boldsymbol{y}^T\boldsymbol{X} \\ &\iff (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y} \iff \hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

where $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ is invertible by the following argument. $\boldsymbol{X}^T\boldsymbol{X}$ must be positive semidefinite. In fact, it is positive definite because $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has full rank; that is, $\text{rank}(\boldsymbol{X}) = p$, so $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{p\times p}$ has rank p (full rank) and is invertible. So $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite (all positive eigenvalues). Then since $\text{Tr}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p) > \text{Tr}(\boldsymbol{X}^T\boldsymbol{X})$, the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ are also all positive, which means the determinant of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ is nonzero, which means it is invertible.

□

Proposition 1.5.0.2 (**DSO 607 Homework Problem**). Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

(a) The asymptotic behavior of the ridge estimator is as follows: as $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}_{\text{ridge}} \to \boldsymbol{0}$, and as $\lambda \to 0$, $\hat{\boldsymbol{\beta}}_{\text{ridge}} \to X^\dagger(X\beta + \epsilon)$.

(b) For any fixed $\lambda > 0$, the probability that each component of the ridge estimator $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ equals 0 is 0.

Proof. (a) Since X is fixed, as $\lambda \to \infty$ we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to (\lambda I_p)^{-1}X^Ty = \frac{1}{\lambda}I_pX^T(X\beta + \epsilon) = \frac{1}{\lambda}(X^TX\beta + X^T\epsilon) \to \boldsymbol{0}$$

where $\boldsymbol{0}$ is a p-dimensional vector of zeroes. As $\lambda \to 0^+$ we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to X^\dagger y = X^\dagger(X\beta + \epsilon)$$

where we substitute the pseudoinverse instead of the inverse because since X^TX is rank deficient, $(X^TX)^{-1}$ does not exist (and because the Moore-Penrose pseudoinverse minimizes the ℓ_2 norm, exactly what the ridge solution will do).

(b) We have

$$\hat{\beta}_{\text{ridge}} = (X^TX+\lambda I_p)^{-1}X^Ty = (X^TX+\lambda I_p)^{-1}X^T(X\beta+\epsilon) = (X^TX+\lambda I_p)^{-1}X^TX\beta+(X^TX+\lambda I_p)^{-1}X^T\epsilon$$

Let e_i be a selection vector, with the ith entry equal to 1 and all other entries equal to 0. Let the ith entry of $\hat{\beta}_{\text{ridge}}$ be $\hat{\beta}^{(i)}_{\text{ridge}} = e_i^T\hat{\beta}_{\text{ridge}}$. We have

$$\Pr(\hat{\beta}^{(i)}_{\text{ridge}} = 0) = \Pr\left(e_i^T[(X^TX + \lambda I_p)^{-1}X^TX\beta + (X^TX + \lambda I_p)^{-1}X^T\epsilon] = 0\right)$$

$$= \Pr\left(e_i^T(X^TX + \lambda I_p)^{-1}X^T\epsilon = -e_i^T(X^TX + \lambda I_p)^{-1}X^TX\beta\right)$$

Since every entry of ϵ is distributed continuously, the probability of it equaling a particular value is 0. Therefore the probability that each component of the ridge estimator equals 0 is 0. (For an intuitive argument as to why this is, see Figure 1.1.)

□

GSBA 604: using SVD $X = UDV^T$, we have

$$\hat{y}_{ridge}(\lambda) = X\hat{\beta}_{ridge}(\lambda) = X(X^TX + \lambda I)^{-1}X^Ty = UDV^T(VD^2V^T + \lambda I)^{-1}VDU^Ty$$
$$= UDV^T(V(D^2 + \lambda I)V^T)^{-1}VDU^Ty = UDV^T[(D^2 + \lambda I)V^T]^{-1}V^TVDU^Ty = UDV^TV(D^2 + \lambda I)^{-1}DU^Ty$$
$$= UD(D^2 + \lambda I)^{-1}DU^Ty = \sum_{j=1}^{n} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^T y \quad (1.7)$$

also,

$$\hat{\beta}_{LS} = \sum_{j=1}^{n} v_j(u_j^T y)$$

so by comparison, what ridge regression is doing is changing the weights of the columns (by weights that are between 0 and 1 for any $\lambda > 0$). Higher d_j's means higher weights. So the directions that have higher variability are shrunk less.

So in the limit of (1.7) as $\lambda \to 0^+$, we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to X^\dagger y = X^\dagger(X\beta + \epsilon) \quad (1.8)$$

where we substitute the pseudoinverse instead of the inverse because since X^TX is rank deficient, $(X^TX)^{-1}$ does not exist (and because the Moore-Penrose pseudoinverse minimizes the ℓ_2 norm, exactly what the ridge solution will do).

$$\hat{y}_{ridge}(\lambda) = X\hat{\beta}_{ridge}(\lambda) \to \sum_{j=1}^{n} u_j u_j^T y = UU^T y,$$

which is the least squares solution in the case that $p \leq n$ and this exists, or (1.8) in the general case.

1.6 Lasso

From KKT theory, the correlation between all selected features and residual will be λ (see the remark in Section 1.6.4 for an explanation why).

Consider the linear regression model $y = X\beta + \epsilon$. If we assume the errors ϵ have a multivariate Gaussian distribution, that is,

$$f_\epsilon(t) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{t^Tt}{2\sigma^2}\right), \quad t = (t_1, \ldots, t_n)^T$$

then the log likelihood is

$$\log(f(t)) = n\log[(2\pi\sigma^2)^{-1/2}] - t^T t/(2\sigma^2)$$

Suppose we want the MLE estimator. When we maximize the log likelihood, we can disregard the first term which does not include t (it is constant). So we seek

$$\arg\max_{\beta\in\mathbb{R}^p} \left\{-t^T t/(2\sigma^2)\right\} = \arg\max_{\beta\in\mathbb{R}^p} \left\{-\|y - X\beta\|_2^2/(2\sigma^2)\right\}$$

which is the same as

$$\arg\min_{\beta\in\mathbb{R}^p} \left\{\|y - X\beta\|_2^2/(2\sigma^2)\right\}$$

We commonly scale this with an n in the denominator to match the empirical risk; note that this does not affect the arguments which minimize the quantity. When the design matrix X multiplied by $n^{-1/2}$ is orthonormal ($X^T X = nI_p$), the penalized least squares reduces to the minimization of

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - X\hat{\beta}\|_2^2 + \frac{1}{2}\|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}$$

where $\hat{\beta} = (X^T X)^{-1} X^T y = nX^T y$ is the OLS estimator. Disregarding the first term which does not contain β, we have a **separable** loss function (we can solve for one parameter at a time):

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}.$$

So we can consider the univariate penalized least squares function

$$\hat{\theta}(z) = \arg\min_{\theta\in\mathbb{R}} \left\{ \frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)) \right\}.$$

Antoniadis and Fan [2001] showed that the PLS estimator $\hat{\theta}$ possesses the following properties:

- *sparsity* if $\min_{t\geq 0}\{t + p'_\lambda(t)\} > 0$;
- *approximate unbiasedness* if $p'_\lambda(t) = 0$ for large t;
- *continuity* if and only if $\arg\min_{t\geq 0}\{t + p'_\lambda(t)\} = 0$. Intuition: if you perturb data a little, the solution should remain similar.

In general, the singularity of the penalty function at the origin (i.e., $p'_\lambda(0+) < 0$) is needed for generating sparsity in variable selection and the concavity is needed to reduce the bias.

To recap: constrained version:

$$\begin{aligned}\hat{\beta}_{\text{lasso}} = &\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \frac{1}{2n}\|y - X\beta\|_2^2 \\ &\text{subject to} \quad \|\beta\|_1 \leq t\end{aligned}$$

Unconstrained version:

$$\hat{\beta}_{\text{lasso}} = \arg\min \left\{ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

Use $1/n$ to rescale RSS due to $\|1\| - 2 = \sqrt{n}$.

Proposition 1.6.0.1 (**Math 541A Homework Problem**)**.** Suppose $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, and for all $1 \leq i \leq n$, there are known vectors $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$. Our observed data are $\boldsymbol{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$. Let $\boldsymbol{X}$ be the $n \times p$ matrix so that the i^{th} row of $\boldsymbol{X}$ is the row vector $x^{(i)}$. Assume that $p \leq n$ and the matrix $\boldsymbol{X}$ has full rank. Let $\lambda > 0$ and consider the quantity

$$\sum_{i=1}^{n} \left(\boldsymbol{y}_i - {x^{(i)}}^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \tag{1.9}$$

Then there exists a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that minimizes this quantity (this $\hat{\boldsymbol{\beta}}$ is known as the LASSO, or least absolute shrinkage and selection operator).

Proof. We can write (1.9) as

$$\begin{aligned}\sum_{i=1}^{n} \left(\boldsymbol{y}_i - {x^{(i)}}^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \end{aligned} \tag{1.10}$$

By Proposition **??**, $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is convex, and by Proposition **??**, $\lambda\|\boldsymbol{\beta}\|_1$ is convex. Therefore by Proposition **??**, (1.10) is convex. Differentiating and setting equal to 0 yields

$$-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + \lambda \left[\text{sgn}(\boldsymbol{\beta}_i) \right] = 0 \tag{1.11}$$

where $\left[\text{sgn}(\boldsymbol{\beta}_i) \right]$ is vector resulting from the sgn function being applied elementwise to $\boldsymbol{\beta}$. Since (1.11) is linear in $\boldsymbol{\beta}$, it has one solution. Since (1.10) is convex, any solution to (1.11) minimizes (1.9).

□

Remark 1. The L_1 penalization term in (1.9) is better at penalizing large entries of $\boldsymbol{\beta}$ (a similar observation applies in the compressed sensing literature). Unfortunately, there is no closed form solution to (1.9) in general. The constrained minimization problem

$$\text{minimize} \sum_{i=1}^{n} (\boldsymbol{y}_i - \langle x^{(i)}, \boldsymbol{\beta} \rangle)^2 \quad \text{subject to} \sum_{i=1}^{n} |\boldsymbol{\beta}_i| \leq 1$$

is morally equivalent to (1.9), but technically Lagrange Multipliers does not apply since the constraint is not differentiable everywhere.

1.6.1 Soft Thresholding

Classical ideas of nonparametric models: kernels (locally constant/linear), splines (smooth basis functions). But wavelets are non-smooth. Why is this beneficial? Some real life functions are non-smooth. (example; image data with noise. There will be non-smooth edges to objects.) Also, the wavelet basis functions are orthonormal (which is closely related to the assumption we made above about the orthonormal design matrix). So when working with wavelets, we have a separable optimization problem. Soft thresholding is something like the lasso idea for wavelets (but before the lasso was developed).

Suppose we wish to recover an unknown function f on $[0, 1]$ from noisy data

$$d_i = f(t_i) + \sigma z_i, \qquad i = 0, \dots, n-1$$

where $t_i = i/n$ and $z_i \sim \mathcal{N}(0, 1)$. The term de-noising is to optimize the mean squared error $n^{-1}\ E\|\hat{f} - f\|_2^2$. Donoho and Johnstone [1994] proposed a soft-thresholding estimator

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where γ is some small number. (So estimator gets shrunk by γ, and if γ is bigger than the original estimator, we set it equal to 0.) They applied this estimator to the coefficients of a wavelet transform of a function measured with noise, then back-transformed to obtain a smooth estimate of the function.

Example 1.6.1. Suppose we have an image in data in the form of $X \in \mathbb{R}^n$. We have a wavelet basis $W \in \mathbb{R}^{n \times n}$ where W is orthonormal. We transform the image into the frequency domain by

$$Wx \to \tilde{x}$$

where $\tilde{x}$ is the frequency domain representation. Then we apply soft-thresholding to $\tilde{x}$ to yield $\tilde{x}^*$, which we hope is de-noised. Finally, we bring the image back into the original domain according to

$$\hat{x} = W^{-1}\tilde{x}^* = W^T\tilde{x}^*.$$

The asymptotic risk of this estimator is

$$[2(\log p) + 1](\sigma^2 + R_{DP})$$

Note that the $2 \log p$ term is related to the result (described informally) below:

Proposition 1.6.1.1. if we have n i.i.d. $\mathcal{N}(0, 1)$ random variables, the maximum of them is near $\sqrt{2 \log n}$ if n is large. (The order is this large with high probability)

Remark 2. In the language of wavelets, sometimes ℓ_0 penalization is called "hard-thresholding."

1.6.2 Lasso theory

Drawbacks of previous techniques that lasso helps with: subset selection is interpretable but computationally intensive and not stable because it is a discrete process (small changes in the data can result in very different models being selected). Ridge regression is a continuous process and more stable, but it does not set any coefficients equal to 0 and hence does not give an easily interpretable model.

In the orthonormal design case $X^TX = nI_p$, the lasso solution can be shown to be the same as soft thresholding:

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where $\gamma \geq 0$ is determined by the condition $\sum_{j=1}^p |\beta_j| = t$.

Geometry: the criterion $\sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2$ equals the quadratic function (plus a constant)

$$(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0).$$

Proof.

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 = \sum_{i=1}^n \left(y_i - X_i\hat{\beta} \right)^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = [\boldsymbol{X}(\beta^0 - \hat{\beta})]^T[\boldsymbol{X}(\beta^0 - \hat{\beta})]$$

$$= (\beta^0 - \hat{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\beta^0 - \hat{\beta})$$

□

The contours (level sets) are therefore elliptical and centered at the OLS estimates. If the constraint region does not have corners, as in ridge regression, zero solutions result with probability zero (see Proposition 1.5.0.2 and Figure 1.1).

Proposition 1.6.2.1 (2018 DSO Statistics Group In-Class Screening Exam, Question 5). Consider the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{1.12}$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$, and $\lambda > 0$.

(a) The following problem is a dual of (1.12):

$$\underset{u \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \quad \text{subject to} \quad \|X^T u\|_\infty \leq \lambda.$$

Also, $\hat{u} = y - X\hat{\beta}$, where $\hat{\beta}$ is a solution of (1.12) and $\hat{u}$ is a solution of the dual.

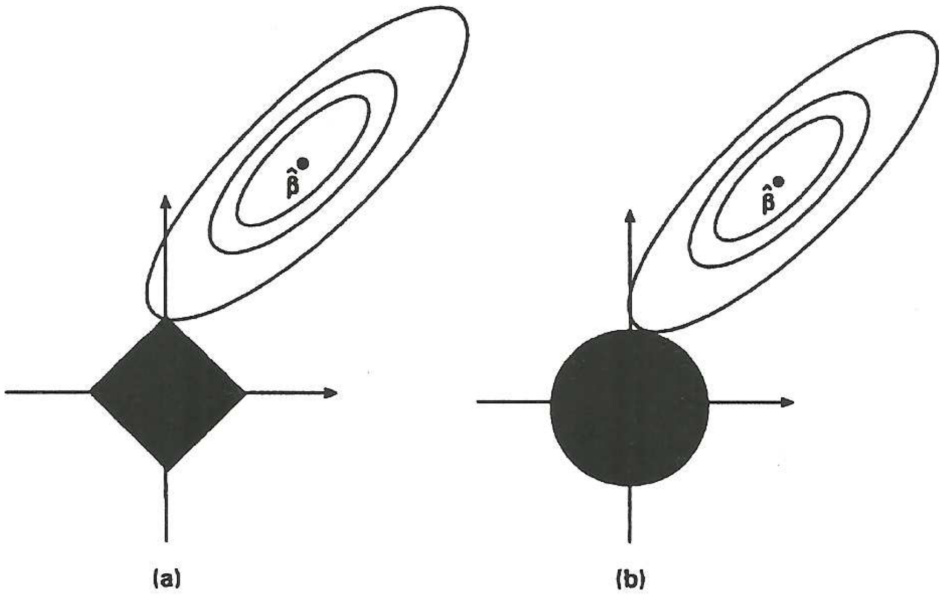

Figure 1.1: Level sets of least squares loss function with feasible sets for (a) lasso and (b) ridge regression in the case of $\beta \in \mathbb{R}^2$.

(b) $\hat{\beta}$ is not necessarily unique, but $\hat{u}$, $\|y - X\hat{\beta}\|_2^2$, and $\|\hat{\beta}\|_1$ are.

(c) Suppose $y = X\beta^* + \epsilon$, and suppose the tuning parameter λ is chosen to satisfy $\lambda \geq \|X^T\epsilon\|_\infty$. Then

(i)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1.$$

(ii)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2.$$

(iii)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|\epsilon\|_2^2 - \lambda\|\beta^*\|_1.$$

Remark 3. We can express the original optimization problem (1.12) as

$$\begin{aligned} \underset{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ \text{subject to} \quad & z = X\beta. \end{aligned} \tag{1.13}$$

We will also refer to another expression of the lasso optimization problem,

$$\begin{aligned} \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2}\|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq t \end{aligned} \tag{1.14}$$

for some $t > 0$.

Before proving the main results, we will show a few simpler results. Whenever $\lambda > 0$, the lasso objective function (1.12) is the Lagrangian of (1.14). We will prove a useful lemma about the relationship between these functions.

Lemma 1.6.2.2. For a given $\lambda > 0$, let $\hat{\beta}$ minimize (1.12). Then there is exactly one $t = \|\hat{\beta}\|_1$ such that any $\hat{\beta}$ minimizing (1.12) also minimizes (1.14).

Proof. This must be true by contradiction. First of all, since the objective function of (1.14) is continuous and the feasible region $\|\beta\|_1 \leq t$ is compact, a minimum of (1.14) is guaranteed to exist. Now suppose $\hat{\beta}$ minimizes (1.12) for a fixed λ, with $\|\hat{\beta}\|_1 = t$, but there is a different solution $\hat{\beta}^*$ that is feasible for (1.14) and achieves a lower value. That is,

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2.$$

and $\|\hat{\beta}^*\|_1 \leq \|\hat{\beta}\|_1 = t$. Since $\lambda > 0$, $\|\hat{\beta}\|_1 < \|\hat{\beta}_{global}\|_1$, where $\hat{\beta}_{global}$ is a global minimum for $\frac{1}{2}\|y - X\hat{\beta}\|_2^2$. Since (1.14) is convex and all global minima lie outside the feasible region, $\hat{\beta}^*$ lies on the boundary; that is, $\|\hat{\beta}^*\|_1 = \|\hat{\beta}\|_1 = t$. But then

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 \iff \frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 + \lambda\|\hat{\beta}^*\|_1 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$$

which contradicts the fact that $\hat{\beta}$ minimizes (1.12).

□

Another useful result follows in a simple way from Lemma 1.6.2.2.

Proposition 1.6.2.3. Let $\mathcal{B}$ be the set of all $\hat{\beta}$ that minimize (1.12) for some fixed $\lambda > 0$. Then for any two $\hat{\beta}_1, \hat{\beta}_2 \in \mathcal{B}$, $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$. That is, $\|\hat{\beta}\|_1$ is unique.

Proof. Suppose $\hat{\beta}_1$ and $\hat{\beta}_2$ both minimize (1.12), and (without loss of generality) $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$. By Lemma 1.6.2.2, these values both minimize (1.14) with $t = \|\hat{\beta}_2\|_1$ (we cannot choose $t = \|\hat{\beta}_1\|_1$ because $\hat{\beta}_1$ is not feasible for that problem). Because the global minimum of (1.14) lies outside the feasible region and (1.14) is convex, all solutions to (1.14) lie on the boundary of the feasible region. But $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$, so $\hat{\beta}_1$ is not on the boundary of the feasible region, contradiction. Therefore $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$ for all solutions $\hat{\beta}_1$, $\hat{\beta}_2$ to (1.12); that is, $\|\hat{\beta}\|_1$ is unique. (See Osborne et al. [2000] for more details.)

□

Now we are ready to prove Proposition 1.6.2.1.

Proof of Proposition 1.6.2.1. (a) The Lagrangian of (1.13) is

$$\mathcal{L}(\beta, z, u) = \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta),$$

so the Lagrange dual function is

$$\inf_{\beta,z}\{\mathcal{L}(x,u)\} = \inf_{\beta,z}\left\{\frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta)\right\}$$

$$= \inf_{\beta,z}\left\{\frac{1}{2}(y-z)^T(y-z) + u^Tz + \lambda\|\beta\|_1 - u^TX\beta\right\}$$

This minimization is separable:

$$= \inf_z\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + u^Tz\right\} + \inf_\beta\left\{\lambda\|\beta\|_1 - u^TX\beta\right\} \tag{1.15}$$

We will handle each part of (1.15) separately. First, the left side:

$$\inf_z\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + u^Tz\right\} = \inf_z\left\{\frac{1}{2}z^Tz + (u-y)^Tz + \frac{1}{2}y^Ty\right\}$$

Since this is a convex quadratic form, differentiate with respect to z and set equal to zero:

$$z + (u-y) = 0 \implies z = y - u \tag{1.16}$$

$$\implies \inf_z\left\{\frac{1}{2}z^Tz + (u-y)^Tz + \frac{1}{2}y^Ty\right\} = \frac{1}{2}(y-u)^T(y-u) + (u-y)^T(y-u) + \frac{1}{2}y^Ty$$

$$= \frac{1}{2}\left(y^Ty - 2u^Ty + u^Tu\right) + 2u^Ty - y^Ty - u^Tu + \frac{1}{2}y^Ty = -\frac{1}{2}u^Tu + u^Ty = \frac{1}{2}y^Ty - \frac{1}{2}y^Ty + u^Ty - \frac{1}{2}u^Tu$$

$$= \frac{1}{2}y^Ty - \frac{1}{2}(y^Ty - 2u^Ty + u^Tu) = \frac{1}{2}y^Ty - \frac{1}{2}(y-u)^T(y-u) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2$$

Next we will minimize the right side of (1.15):

$$\inf_\beta\left\{\lambda\|\beta\|_1 - u^TX\beta\right\} = \inf_\beta\left\{\lambda\sum_{i=1}^p|\beta_i| - \sum_{i=1}^p\left[u^TX\right]_i\beta_i\right\} = \inf_\beta\left\{\sum_{i=1}^p\left(\lambda|\beta_i| - \left[u^TX\right]_i\beta_i\right)\right\}$$

$$= \inf_\beta\left\{\sum_{i=1}^p\left(\text{sgn}(\beta_i)\lambda - \left[u^TX\right]_i\right)\beta_i\right\} = \sum_{i=1}^p\inf_{\beta_i}\left\{\left(\text{sgn}(\beta_i)\lambda - \left[u^TX\right]_i\right)\beta_i\right\}.$$

Notice that when β_i is negative, if $\left(\text{sgn}(\beta_i)\lambda - \left[u^TX\right]_i\right) = -\left(\lambda + \left[u^TX\right]_i\right)$ is positive there is no lower bound on the quantity we are minimizing; otherwise, when β_i is negative the infimum is 0. When β_i is positive, if $\left(\text{sgn}(\beta_i)\lambda - \left[u^TX\right]_i\right) = \left(\lambda - \left[u^TX\right]_i\right)$ is negative there is no lower bound on the quantity we are minimizing; otherwise, when β_i is negative the infimum is 0. That is, the only dual feasible points satisfy for all i

$$-\left(\lambda + \left[u^TX\right]_i\right) \le 0, \qquad \lambda - \left[u^TX\right]_i \ge 0 \iff \left[u^TX\right]_i \ge -\lambda, \qquad \left[u^TX\right]_i \le \lambda$$

which is equivalent to the condition

$$\|u^TX\|_\infty \le \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta,z}\{\mathcal{L}(x,u)\} = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 \tag{1.17}$$

subject to the constraint $\|u^T X\|_\infty \leq \lambda$. This quantity represents a lower bound on the minimum value of the original optimization problem for all $u \in \mathbb{R}^p$. The dual problem is to find the best lower bound by maximizing over u; that is, the dual problem is

$$\begin{aligned} &\underset{u\in\mathbb{R}^p}{\text{maximize}} && \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 \\ &\text{subject to} && \|u^T X\|_\infty \leq \lambda. \end{aligned} \tag{1.18}$$

Lastly, suppose $\hat{\beta}$ and $\hat{u}$ satisfy

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad ,$$

$$\hat{u} = \begin{array}{ll} \underset{u\in\mathbb{R}^p}{\arg\max} & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda \end{array} = \begin{array}{ll} \underset{u\in\mathbb{R}^p}{\arg\min} & -\frac{1}{2}\|y\|_2^2 + \frac{1}{2}\|y-u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda \end{array}$$

Then since (1.16) is a requirement for dual feasibility of u and strong duality applies, we have $\hat{u} = y - X\hat{\beta}$.

Remark 4. Note that we could also find the arguments maximizing (1.18) by

$$\begin{array}{ll} \underset{u\in\mathbb{R}^p}{\arg\min} & -\frac{1}{2}\|y\|_2^2 + \frac{1}{2}\|y-u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda. \end{array} = \begin{array}{ll} \underset{u\in\mathbb{R}^p}{\arg\min} & \frac{1}{2}\|y-u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda. \end{array}$$

where the first step follows from the fact that arguments that maximize a function are the same as the arguments that minimize the negative of a function, and the second step follows from the fact that the $-\frac{1}{2}\|y\|_2^2$ term does not include u. Therefore we see that the residual vector u from a lasso fit can be thought of as the projection of y onto the convex polyhedron $C \subset \mathbb{R}^n$ defined by $C := \{u : \|X^T u\|_\infty \leq \lambda\}$.

Another way of saying this is that the lasso estimate $\hat{y} = X\hat{\beta}_{\text{lasso}}$ itself is the residual from projecting y onto C; that is,

$$X\hat{\beta}_{\text{lasso}} = (I - P_C)y,$$

where P_C is the operator projecting y onto C.

(b) (i) **Not necessarily unique.** Per Tibshirani [2013], if $\text{rank}(X) < p$, the lasso solution is not necessarily unique. Intuitively, this is because the columns of X are linearly dependent, so there may exist more than one linear combination of the columns that minimizes (1.12). **Jacob's suggestion: counterexample. X is two columns that are equal; then convex combinations of two solutions are equal as long as same sign (can't be opposite sign because then ℓ_1 could be smaller by setting one equal to 0.**

(ii) **Necessarily unique.** The dual problem (1.18) is strictly concave, so the value $\hat{u}$ that maximizes it is unique.

(iii) **Necessarily unique** (except in the trivial case $\lambda = 0$). Per part 5(b)(iv), $\|\hat{\beta}\|_1$ is unique. (1.12) is convex, so the minimum $\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$ is unique. Therefore $\|y - X\hat{\beta}\|_2^2$ must be unique.
Jacob's solution: Since $\hat{u}$ is unique and by (1.16) $\hat{u} = y - X\hat{\beta}$, we must have that $\|\hat{u}\| = \|y - X\hat{\beta}\|$ is unique.

(iv) **Necessarily unique** (except in the trivial case $\lambda = 0$). This is immediate from Proposition 1.6.2.3.

(c) (i) Since β^* is clearly feasible for (1.12) and $\hat{\beta}$ achieves the minimum, we have

$$\frac{1}{2}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 \geq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \iff \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1$$

(ii) We know that the expression in the dual problem (1.18) is a lower bound for the solution of the primal problem (1.12) for any u feasible for (1.18) (that is, any u satisfying $\|u^T X\|_\infty \leq \lambda$). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2.$$

Since by assumption $\lambda \geq \|X^T\epsilon\|_\infty$, ϵ is feasible for (1.18). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \epsilon\|_2^2 = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \tag{1.19}$$

as desired.

(iii) We can rewrite the right side of (1.19) as

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|X\beta^*\|_2^2 + \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^* - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^*. \tag{1.20}$$

By assumption, we have

$$\lambda \geq \|X^T\epsilon\|_\infty \iff \lambda\mathbf{1} - X^T\epsilon \succeq 0 \implies \lambda\mathbf{1}\beta^* - X^T\epsilon\beta^* \succeq 0$$

$$\iff -\lambda\|\beta^*\|_1 \leq \epsilon^T X\beta^* \leq \lambda\|\beta^*\|_1.$$

By Hölder's Inequality, we have for any two vectors $u, v \in \mathbb{R}^n$, $|u^T v| \leq \|u\|_\infty\|v\|_1$. Therefore

$$|\epsilon^T X\beta^*| = |(X^T\epsilon)^T\beta^*| \leq \|X^T\epsilon\|_\infty\|\beta^*\|_1 \leq \lambda\|\beta^*\|_1$$

where the last step used the assumption $\|X^T\epsilon\|_\infty \leq \lambda$. So we have

$$\frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^*.$$

Substituting in to (1.19), using the identity in (1.20), and using the result from part 5(c)(iii) yields

$$\frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^* = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1$$

as desired.

(iv) We see from parts (i) and (iii) that

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 - \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 + \lambda\|\beta^*\|_1$$

$$\iff \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1 - \frac{2}{n}\lambda\|\beta^*\|_1 \leq \frac{1}{n}\|\epsilon\|_2^2 \leq \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1 + \frac{2}{n}\lambda\|\beta^*\|_1$$

that is, we can lower bound and upper bound $\frac{1}{n}\|\epsilon\|_2^2$ by taking the quantity $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1$ and adding or subtracting $\frac{2}{n}\lambda\|\beta^*\|_1$. Therefore it seems that the quantity in the middle of this interval, $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1$, is a reasonable estimator for $\sigma^2 = \mathbb{E}\left[n^{-1}\|\epsilon\|_2^2\right]$.

□

1.6.3 Non-Negative Garotte

This idea inspired the lasso. Proposed by Breiman [1995]. It minimizes

$$\sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{p} c_j\hat{\beta}_j^o x_{ij}\right)^2 \text{ subject to } c_j \geq 0, \sum_{j=1}^{p} c_j \leq t$$

It starts with OLS estimates and shrinks them by non-negative factors whose sum is constrained. It depends on both the sign and magnitude of OLS estimates. In contrast, lasso avoids the explicit use of OLS estimates.

1.6.4 LARS—Preliminaries and Intuition

Intuition: the algorithm takes steps from a model where all coefficients are 0 to the biggest model (the unpenalized OLS model). Covariates are considered from the highest correlation with y to the least. (The variable most highly correlated with y is the one at the "least angle" from y.) Recall the original definition of the lasso estimator:

$$\hat{\beta}_{lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2n}\|y - X\beta\|_2^2\right\} \text{ subject to } \|\beta\|_1 \leq t \tag{1.21}$$

The more common version now:

$$\hat{\beta}_{lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} \tag{1.22}$$

One form can be changed to the other by applying Lagrangians[1]. Have to be careful because this is a convex program (quadratic with "linear" constraint—use a slack variable).

[1] However, the correspondence between t and λ is **not** one-to-one. Because with $t = \infty$, $\lambda = 0$. But a slightly smaller t would result in the same solution.

Taking the gradient of the loss function in (1.22) yields

$$\nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) = \nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2\right) + \lambda\nabla\big(\|\beta\|_1\big)$$

$$= -\frac{1}{n}X^T(y - X\beta) + \lambda\nabla\big(\|\beta\|_1\big) \tag{1.23}$$

We set this equal to zero. If the first term equals 0, the residual has to equal 0. For the second part to equal zero, we have to account for the fact that the gradient doesn't exist at 0. In the one-dimensional case $g(t) = |t|$, we have

$$g'(t) = \begin{cases} -1 & t < 0 \\ 1 & t > 0 \end{cases}$$

but it doesn't exist at 0. Instead of using the gradient, we will use ∂, the subdifferential, which is the set of all subgradients. We have a solution if 0 is in the subdifferential. We can rewrite (1.23) using the subdifferential instead of the gradient:

$$\partial\left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) = \nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2\right) + \lambda\partial\big(\|\beta\|_1\big) = -\frac{1}{n}X^T(y - X\beta) + \lambda\partial\big(\|\beta\|_1\big)$$

Then rather than setting the gradient equal to 0, our condition is

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda\partial\big(\|\beta\|_1\big)$$

Note that

$$\partial g(t) = \begin{cases} -1 & t < 0 \\ [-1, 1] & t = 0 \\ 1 & t > 0 \end{cases} = \begin{cases} \text{sgn}(t) & t \neq 0 \\ [-1, 1] & t = 0 \end{cases}$$

so we have

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda \cdot \left[\begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases}\right] \tag{1.24}$$

where

$$\left[\begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases}\right] \in \mathbb{R}^p$$

is a vector with each entry as specified.

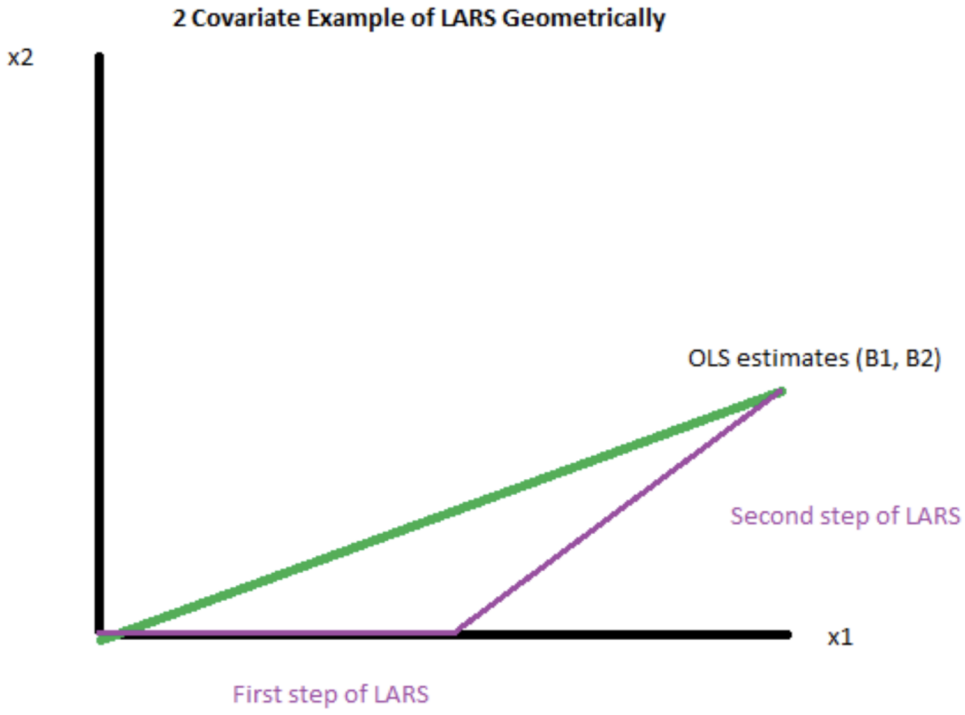

Figure 1.2: LARS figure in 2d case.

Remark 5. (1) Examining the jth component of the separable equation (1.24), if $\beta_j \neq 0$, we have

$$0 = -\frac{1}{n} X_j^T (y - X\beta) + \lambda \cdot \text{sgn}(\beta_j) \iff \frac{1}{n} X_j^T (y - X\beta) = \lambda \cdot \text{sgn}(\beta_j)$$

Note that the left side contains the correlation between X_j and $e = y - X\beta$, the residual vector. **So if lasso chooses k variables, all k of them will have the same correlation with the residual (λ).**

(2) If $\beta_j = 0$, we have

$$0 \in -\frac{1}{n} X^T (y - X\beta) + \lambda \cdot [-1, 1] \iff \left| \frac{1}{n} X^T (y - X\beta) \right| \leq \lambda$$

So for unselected features, the (absolute) correlation should be bounded by λ.

These two conditions relate to the KKT conditions (first order conditions).

So if we start with λ very large and gradually decrease it, we will let in as the first feature the one that is most highly correlated with y—that is, the feature with the *least angle* between it and y.

1.6.5 LARS

In Figure 1.2, note that we choose feature X_1 first because it has the highest correlation with y. As the coefficient on X_1 increases, the correlation between X_1 and the residual with y decreases, while the corre-

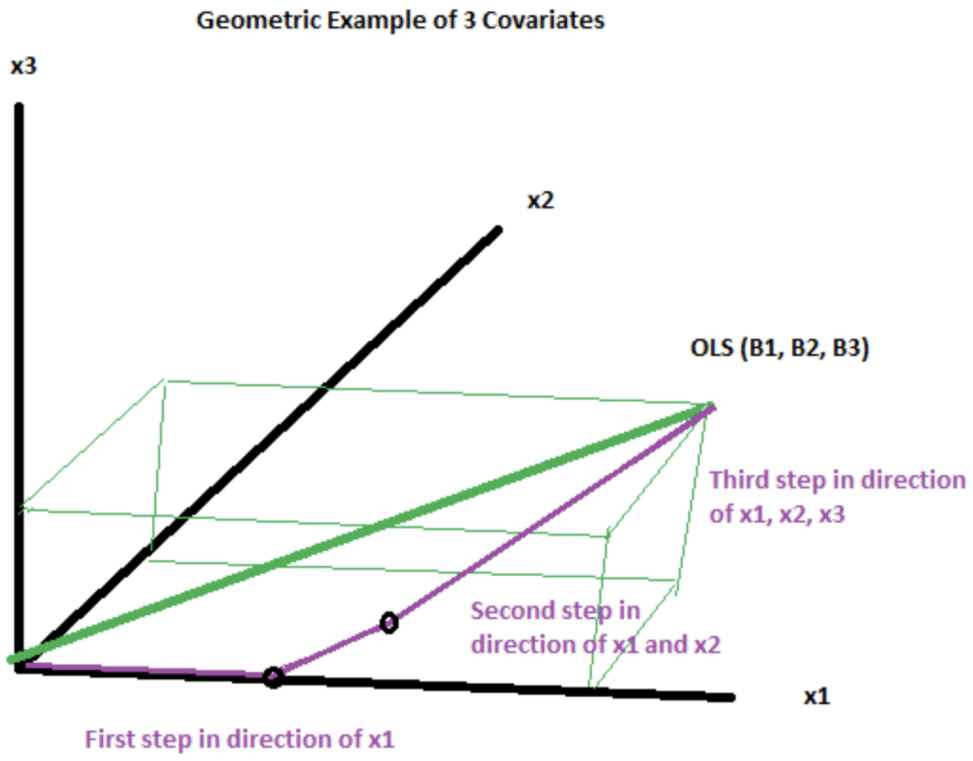

Figure 1.3: LARS figure in 3d case.

lation between X_2 and the residual remains constant (**increases?**). When the correlation between X_1 and the residual becomes equal to the correlation between X_2 and the residual, X_2 enters the lasso path.

Remark 6. Just like in lasso, in LARS the correlation between all included features and the residual are equal (see the remark in Section 1.6.4). However, LARS is a stepwise procedure—once we add a feature, it stays in the model. In the lasso, features can be dropped later in the path after they are selected—whenever β_j becomes 0, it is dropped from the current active set. A feature's sign cannot change in lasso—it is not possible. If we modify the LARS algorithm to have this property ("lasso modification"), then the result is the lasso estimator.

The LARS algorithm for lasso has order $\mathcal{O}(np \cdot \min\{n, p\})$. In particular, if $p > n$ it has order $\mathcal{O}(n^2 p)$.

1.7 Loss Functions

Asymmetric loss: may have asymmetry between overestimation and underestimation. For example: in a supply chain, estimate inventory level $y_1, \ldots, y_n$. Let $y_i = x_i'\beta + \epsilon_i$. Underestimating is really bad because then you don't have enough product for customers and they might not come back; overestimating not so bad. Then our loss function:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (1-\tau) \underbrace{(y_i - x_i'\beta)_+}_{\text{under-estimation}} + \tau \underbrace{(x_i'\beta - y_i)_+}_{\text{over-estimation}}$$

you could choose $\tau = 0.1$ for a 9 to 1 ratio of loss (i.e., underestimation is 9 times as costly as overestimation).

Theorem 1.7.0.1 (Loss: quadratic). Let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}X^2 < \infty$. Then $\mathbb{E}(X-t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbb{E}X$.

Proof. We seek

$$\arg\min_t \mathbb{E}(X-t)^2 = \arg\min_t \left[\mathbb{E}(X^2) - 2t\mathbb{E}(X) + t^2\right] = \arg\min_t \left[t^2 - 2t\mathbb{E}(X)\right]$$

where the last step follows because $\mathbb{E}(X^2)$ is independent of t. This expression is quadratic in t. Differentiating with respect to t and setting equal to 0, we have

$$2t - 2\mathbb{E}(X) = 0 \implies \boxed{\arg\min_t \mathbb{E}(X-t)^2 = \mathbb{E}(X)}$$

□

Huber loss: combines benefits of squared loss (unbiasedness) with MAD loss (robust).

$$L_\delta^H(y,\hat{y}) = (y-\hat{y})^2 \cdot I\left\{|y-\hat{y}| \le \delta\right\} + |y-\hat{y}| \cdot I\left\{|y-\hat{y}| > \delta\right\}.$$

Only downside: not that easy to estimate (loss function is not differentiable). Instance of **M-estimation** (more generalized than regression):

$$\hat{\beta}_M = \arg\min_{\beta,\sigma} \sum_{i=1}^n \rho_M\left(\frac{y_i - x_i'\beta}{\sigma}\right)$$

where ρ_M is a loss function. Suppose $y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i$. Then differentiating the loss function with respect to β yields

$$\frac{\partial}{\partial \beta_k}\left[\sum_{i=1}^n \rho_M\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] = 0 \qquad \text{for } k \in \{1,\dots,p\}.$$

$$= \sum_{i=1}^n \rho'(y_i - x_i'\beta) \cdot \frac{(-x_{ij})}{\sigma}$$

Let $u_i := (y_i - x_i'\beta)/\hat{\sigma}$.

$$\implies \sum_{i=1}^n \underbrace{\frac{\rho'(u_i)}{u_i}}_{\text{weight}} \cdot x_{ij}(y_i - x_i'\beta) = 0$$

Compare to least squares: normal equations $X^T(y - X^T\beta_{OLS}) = 0$, or

$$\sum_{i=1}^{n} x_{ij}(y_i - x_i'\beta) = 0$$

Now we have this extra weighting term. Algorithm for computing:

1. Use $\hat{\beta}_{OLS}$ as the initial solution; then compute

$$u_i^{(0)} = \frac{y_i' - x_i'\hat{\beta}_{OLS}}{\hat{\sigma}_{OLS}}$$

2. With $w_i^{(0)} = \rho'(u_i)/u_i$, compute

$$\hat{\beta}_{LS}^{(1)} = (X^T W X)^{-1} X^T W y.$$

3. Update weights with $\hat{\beta}_{LS}^{(1)}$.

4. Repeat until convergence. (Convergence is a little tricky but should work.)

For absolute deviation loss, see Section **??** on quantile regression.

1.7.1 Feature Selection properties

Model selection consistency: $\Pr(\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)) \to 1$.

Oracle property: model selection consistency, asymptotic efficiency as efficient as if true model were known ("efficiency" having to do with the variance given n).

Definition 1.7.1 (**Oracle property**)**.** Let β^0 denote the true parameter vector for data generated from a linear model. Let S_0 be the true support; that is, $S_0 = \{j : \beta_j^0 \neq 0, j = 1, \ldots, p\}$. Denote $\hat{\beta}(\delta)$ the coefficient estimator for fitting procedure δ. We call δ an **oracle procedure** if $\hat{\beta}(\delta)$ asymptotically has the following properties:

- Identifies the right subset model (consistency): $\{j : \hat{\beta}_j \neq 0\} = S_0$.
- Has the optimal estimation rate: $\sqrt{n}(\hat{\beta}(\delta)_{S_0} - \beta_{S_0}^0) \xrightarrow{d} \mathcal{N}(0, \Sigma_0)$ where Σ_0 is the covariance matrix knowing the true subset model.

The lasso problem is convex but not necessarily strictly convex if $p > n$. That is, there is some flat region, so the minimizer may not be unique. Consider the KKT conditions from convex optimization:

$$g(\beta) = \arg\min \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} = \arg\min \{f_1(\beta) + f_2(\beta)\}$$

Then $\hat{\beta}$ is a lasso solution if and only if 0 is in the subdifferential of $g(\hat{\beta})$. Note that

$$\partial g(\hat{\beta}) = \nabla f_1 + \partial f_2 = \frac{1}{n} X^T (X\beta - y) + \lambda \begin{bmatrix} \vdots \\ \partial |\beta_j| \\ \vdots \end{bmatrix} = \frac{1}{n} X^T (X\beta - y) + \lambda \begin{bmatrix} \vdots \\ \begin{cases} \text{sgn}(\beta_j) & \beta_j \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \\ \vdots \end{bmatrix}$$

Now assume $\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)$ (that is, assume lasso recovers the correct support). Suppose the first s features are nonzero and consider one of them (so we know that we should have $\hat{\beta}_j \neq 0$):

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n} X^T (X\beta - y) \right]_j + \lambda \, \text{sgn}(\hat{\beta}_j)$$

Therefore

$$\frac{1}{n} X_A^T (X\hat{\beta} - y) + \lambda \, \text{sgn}(\hat{\beta}_j) = 0 \tag{1.25}$$

where X_A is a submatrix of X containing the columns corresponding to the features in the true support, is our first condition. Next, consider what happens for $j > s$ (features not in the true support). We have

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n} X^T (X\beta - y) \right] + \lambda [-1, 1]$$

$$\implies \left\| \frac{1}{n} X_{A^C}^T (X\hat{\beta} - y) \right\|_\infty \leq \lambda \tag{1.26}$$

where X_{A^C} is a submatrix of X containing the columns corresponding to the features not in the true support, is our boundary condition. Recall the true model

$$y = X\beta_0 + \epsilon$$

and consider the case $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ where X_1 are the features in the true model and X_2 are noise features; that is, $\beta_0 = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}$. Then we are assuming

$$\hat{\beta}_{\text{lasso}} = \begin{bmatrix} \hat{\beta}_1 \\ 0 \end{bmatrix}.$$

We have from (1.25)

$$0 = \frac{1}{n} X_1^T (X\hat{\beta} - y) + \lambda \, \text{sgn}(\hat{\beta}_1) = \frac{1}{n} X_1^T (X_1 \hat{\beta}_1 - X_1 \beta_1 - \epsilon) + \lambda \, \text{sgn}(\hat{\beta}_1)$$

$$\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \, \text{sgn}(\hat{\beta}_1)$$

Let's assume that $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta_0)$ (sign consistency).

$$\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1)$$

which is linear in $\hat{\beta}$. Solving, we have

$$\iff \hat{\beta}_1 - \beta_1 = (X_1^T X_1)^{-1} \big(X_1^T \epsilon - n\lambda \, \text{sgn}(\beta_1) \big) \iff \hat{\beta}_1 = \beta_1 + (n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) \quad (1.27)$$

Looking at the second (boundary) condition (1.26), we have

$$\left\| \frac{1}{n} X_2^T (X\hat{\beta} - y) \right\|_\infty \leq \lambda. \quad (1.28)$$

Consider that

$$X\hat{\beta} - y = X_1 \hat{\beta}_1 - X_1 \beta_1 - \epsilon = X_1 (\hat{\beta}_1 - \beta_1) - \epsilon$$

Substituting in the result from (1.27) yields

$$X\hat{\beta} - y = X_1 \big[(n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) \big] - \epsilon$$

which when we plug into (1.28) yields

$$\left\| \frac{1}{n} X_2^T \left[X_1 (n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) - \epsilon \right] \right\|_\infty \leq \lambda.$$

$$\iff \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) - \frac{1}{n} X_2^T \epsilon \right\|_\infty \leq \lambda.$$

Using the Triangle Inequality, we have

$$\left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) - \frac{1}{n} X_2^T \epsilon \right\|_\infty$$

$$\leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \big(n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \big) \right\|_\infty + \left\| \frac{1}{n} X_2^T \epsilon \right\|_\infty$$

$$\leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \right\|_\infty \cdot \left\| n^{-1} X_1^T \epsilon - \lambda \, \text{sgn}(\beta_1) \right\|_\infty + \left\| \frac{1}{n} X_2^T \epsilon \right\|_\infty \quad (1.29)$$

Assume that the jth column of X has L_2 norm $n^{1/2}$ (as it would if all entries equaled 1). We have

$$\left\|n^{-1}X_1^T\epsilon\right\|_\infty \leq \lambda/2, \qquad \left\|n^{-1}X_2^T\epsilon\right\|_\infty \leq \lambda/2$$

$$\left\|n^{-1}X^T\epsilon\right\|_\infty \leq \lambda/2 \text{ with large probability}$$

Recall that $\lambda = \sigma\sqrt{\frac{c\log p}{n}}$ for some $c > 2$. Then we have (continuing from (1.29)), and using $\left\|n^{-1}X_2^T\epsilon\right\| \leq \lambda/2$,

$$\leq \left\|n^{-1}X_1^T\epsilon\right\|_\infty + \|\lambda\,\mathrm{sgn}(\beta_1)\|_\infty$$

$$\left\|n^{-1}X_2^TX_1(n^{-1}X_1^TX_1)^{-1}\right\|_\infty \cdot \underbrace{\|\cdot\|_\infty}_{3/2\lambda} + \underbrace{\|\cdot\|_\infty}_{\lambda/2} \leq \lambda$$

$$\left\| \underbrace{n^{-1}X_2^TX_1}_{\text{corr. between noise and true}} \quad (\quad \underbrace{n^{-1}X_1^TX_1}_{\text{sample covariance matrix}} \quad)^{-1} \right\|_\infty \leq 1/3 \tag{1.30}$$

It turns out we're fine as long as it's less than or equal to 1. This is known as the **irrepresentable condition**. Note that the sample covariance matrix is the same as the sample correlation since the columns are standardized. So this is the correlation between the true variables. Note that this matrix has dimension $(p-s)\times s$ where s is the dimension of the true support. Note that

$$n^{-1}X_2^TX_1(n^{-1}X_1^TX_1)^{-1} = (X_2^TX_1(X_1^TX_1)^{-1}]^T = (X_1^TX_1)^TX_1^TX_2$$

which is ordinary least squares for regressing X_2 on X_1. In the end, the irrepresentable condition says the correlation between the noise and true variables can't be too high.

1.8 Dantzig Selector

Dantzig selector:

$$\begin{aligned}\hat{\beta}_{\text{Dantzig}} = &\arg\min_{\beta\in\mathbb{R}^p} && \|\beta\|_1 \\ &\text{subject to} && \|n^{-1}X^T(y-X\beta)\|_\infty \leq \lambda\end{aligned}$$

Can be recast as a linear program:

$$\begin{aligned}\hat{\beta}_{\text{Dantzig}} = &\arg\min_{u\in\mathbb{R}^p} && \sum_{i=1}^p u_i \\ &\text{subject to} && -u \leq \beta \leq u \\ & && -\lambda_p\sigma\mathbf{1} \leq n^{-1}X^T(y-X\beta) \leq \lambda_p\sigma\mathbf{1}\end{aligned} \tag{1.31}$$

where $|u|$ denotes the absolute value of u componentwise. (This is a benefit because linear programming is easy to use and very popular in industry and other applications.) Note that $n^{-1}X^t(y-X\beta)$ corresponds to the correlations between the residuals and the design matrix. Recall that in OLS this correlation is 0—the design matrix is orthogonal to the residuals. In the Dantzig selector we relax this, bounding the L_∞ norm by λ. Recall that the gradient of the log-likelihood is the **score function**, in this case $n^{-1}X6t(y - X\beta)$. For example, the score equation in linear regression is $n^{-1}X^Ty = n^{-1}X^TX\beta$. Note:

$$\nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2\right) = \frac{1}{n}X^T(X\beta - y)$$

Note for Theorem 1: in original paper, assumed columns had L_2 norm 1, resulting in $\lambda_p = \sqrt{2\log p}$. We are instead assuming each column has L_2 norm $\sqrt{n}$, which results in $\lambda = \sigma \cdot \sqrt{\frac{c\log p}{n}}$. Intuition of $\log p$ term:

By a theorem in James et al. [2009], the lasso and Dantzig selector estimates equal each other under certain conditions:

Theorem 1.8.0.1. Let I_l be the support of the lasso estimate $\hat{\beta}_{\text{lasso}}$. Let $\boldsymbol{X}_L$ be the $n \times |I_L|$ matrix constructed by taking $\boldsymbol{X}_{I_L}$ and multiplying its columns by the signs of the corresponding coefficients in $\hat{\beta}_{\text{lasso}}$. Suppose that $\lambda_{\text{lasso}} = \lambda_{\text{Dantzig}}$. Then $\hat{\beta}_{\text{lasso}} = \hat{\beta}_{\text{Dantzig}}$ if $\boldsymbol{X}_L$ has full rank and

$$\boldsymbol{u} = \left(\boldsymbol{X}_L^T\boldsymbol{X}_L\right)^{-1}\mathbf{1} \succeq 0 \text{ and } \|\boldsymbol{X}^T\boldsymbol{X}_L\boldsymbol{u}\|_\infty \leq 1$$

where $\mathbf{1}$ is an $|I_L|$-vector of ones and the vector inequality is understood componentwise.

Corollary 1.8.0.1.1. If $\boldsymbol{X}$ is orthonormal ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$), then the entire lasso and Dantzig selector coefficient paths are identical.

Proof. For each index set $\boldsymbol{I}$, $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_{|\boldsymbol{I}|}$, so clearly both of the conditions of Theorem 1.8.0.1 are satisfied.

□

The entire paths can be identical under another condition presented in the same paper.

Theorem 1.8.0.2. Suppose that all pairwise correlations between the columns of $\boldsymbol{X}$ are equal to the same value ρ where $0 \leq \rho < 1$. Then the entire Lasso and Dantzig selector coefficient paths are identical. In addition, when $p = 2$, the same holds for every $\rho \in (-1, 1)$.

1.9 Coordinate Descent

Start with β_1 varying and all other βs fixed. Optimize β_1. Then cycle through each β_j, run until convergence.

1.10 Total Variational Distance

1.11 Non-parametric regression

One example: LOESS

Another: GAM

1.11.1 Generalized additive models

Suppose $y_i = f(x_i) + \epsilon_i$. Suppose $0 \leq x_1 \leq \ldots \leq x_n \leq 1$. Then express

$$f(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x)$$

where $\int_{\mathbb{R}} f^2(x)\ dx < \infty$ and the ϕ_i form an orthonormal basis. Assume sparsity:

$$f(x) \approx \sum_{i=1}^{p} \theta_i \phi_i(x).$$

So

$$y_i = \sum_{i=1}^{p} \theta_i \phi_i(x) + \epsilon_i, \qquad i \in \{1, \ldots, n\}$$

and if f is smooth then the coefficient basis is sparse, so θ is sparse, so

$$\|\theta\|_p = \left[\sum_{i=1}^{n} |\theta_i|^p \right]^{1/p}$$

is small.

Can choose a wavelet basis.

1.12 Mixture regression

Suppose

$$y_i \sim \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(X_i' \beta_k, \sigma_k^2 I\right).$$

Mixture modeling is very useful. Typically estimated by expectation maximization (see Section **??**). One example: spatial analysis. See Figure 1.4. In the black region, slopes and standard deviations will be much higher; outside, will be much lower.

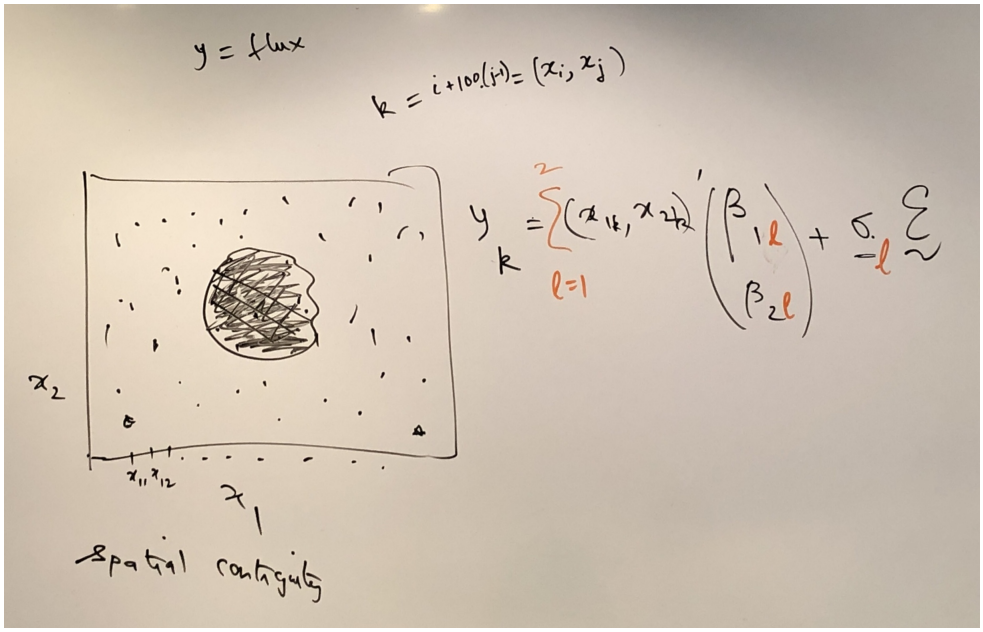

Figure 1.4: Mixture model example.

1.13 Missing observations

1.14 Generalized linear models

Exponential family: binomial, Poisson, Beta, negative binomial, etc. See Section **??** for more information on exponential families. The density is

$$g_\eta(y) = c(y) \exp\{\eta y - \psi(\eta)\}$$

where $c(y)$ is the **carrier density**, η is the **natural parameter**, y is a **sufficient statistic**, and $\psi(\cdot)$ is the **normalizing function** (or **cumulant generating function**), chosen so that it is a proper density: that is, choose $\psi(\eta)$ such that

$$\int_{\mathbb{R}} g_\eta(y)\ dy = 1,$$

so

$$e^{-\psi(y)} = \left(\int_{\mathbb{R}} c(y) e^{\eta y}\ dy \right)^{-1}$$

The density is simplest to express in the natural parameter.

Skewness: third centered moment.

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^3}{\text{Var}(Y)^{3/2}}$$

kurtosis: 4th centered moment

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^4}{\text{Var}(Y)}$$

get excess kurtosis by subtracting 3 (which is Gaussian kurtosis).

In general: cumulant, either centered or raw.

$$K_\gamma = \mathbb{E}(Y - \mathbb{E}(Y))^\gamma, \qquad K_\gamma = \mathbb{E}(Y^\gamma)$$

Cumulant generating function:

$$\psi(\eta) - \psi(\eta_0) = \sum_{\gamma=1}^{\infty} K_\gamma \frac{(\eta - \eta_0)^r}{r!}$$

if you know all the cumulants, you know the distribution exactly.

1. **Gaussian variables:** Suppose $y_1, y_2, \dots y_n$ are i.i.d. $\mathcal{N}(\mu, 1)$.
2. **Gaussian response:** Suppose (y_i, x_i) are i.i.d. pairs in $\mathbb{R}^{p+1}$ with $y_i \sim \mathcal{N}(x_i'\beta, 1)$.
3. **Exponential family variables:** Suppose $Y_1, \dots, Y_n$ are i.i.d. in an exponential family. Then the density of each is $g_\eta(y) = e^{\eta y 0 \psi(\eta)} c(y)$ and the joint density is

$$\prod_{i=1}^{n} g_\eta(y_i) = \exp\left(\eta \sum_{i=1}^{n} y_i - n\psi(\eta)\right) \cdot \prod_{i=1}^{n} c(y_i). \tag{1.32}$$

   Then the density of $\overline{y}$ is simply

$$\exp\left\{n\eta\overline{y} - n\psi(\eta)\right\}$$

   now the natural parameter is $n\eta$ and the normalizing function is $n\psi(\eta)$. So

$$\overline{Y} \sim \left[n\frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}, n\frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2}\right].$$

   The log likelihood is the log of (1.32):

$$\log L(\eta) = \ell(\eta) = \sum_{i=1}^{n} \eta y_i - n\psi(\eta) + \sum_{i=1}^{n} \log c(y_i).$$

The maximum likelihood estimate solves

$$\hat{\eta}_{MLE} := \arg\max_{\eta} \ell(\eta)$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) = \sum_{i=1}^{\eta} y_i - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta), \qquad \frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\ell(\eta) = -n\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\psi(\eta) = -n\mathrm{Var}(\eta) < 0$$

so since this function is concave down it has a unique maximum at

$$\sum_{i=1}^{n} y_i - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\hat{\eta}_{MLE}) \implies \psi(\hat{\eta}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \implies \hat{\mu}_{MLE} = \overline{y}$$

That is, the MLE of the mean parameter in an exponential family is the sample mean. Due to the equivariance property of the maximum likelihood estimator, the MLE for any function of the mean parameter is simply the function of the MLE for the mean parameter (see Proposition **??**).

We call the derivative $\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)$ the **score function**. If $\xi = h(\eta)$, then by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\ell(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) \Big/ \frac{\mathrm{d}\xi}{\mathrm{d}\eta}$$

For the mean parameter in particular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\ell(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) \Big/ \frac{\mathrm{d}\mu}{\mathrm{d}\eta} = n(\overline{y} - \mu)/\mathrm{Var}_{\eta}(Y) \tag{1.33}$$

(since $\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) = n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)$ and $\mathrm{Var}_{\eta}(Y) = \frac{\mathrm{d}\mu}{\mathrm{d}\eta}$). Also, since the expectation of the score function $\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)) = 0$, we hace

$$\mathrm{Var}\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right] - \left[\mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)\right]^2 = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right]$$

We call this the **Fisher information:**

$$i_{\eta}^{(n)}(\xi) := \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right]. \tag{1.34}$$

(See also Definiton **??** and Section **??** for more on this.) In the case of the mean, we can write this as

$$i_{\eta}^{(n)}(\eta) = \mathbb{E}\left[\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right)^2\right] = \mathrm{Var}\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right) = \mathrm{Var}(n\overline{y}) = n^2\mathrm{Var}(\overline{y}) = n^2\mathrm{Var}_{\eta}(y)/n = n\mathrm{Var}_{\eta}(y)$$

Then substituting (1.33) into (1.34) we have

$$i_{\eta}^{(n)}(\mu) = \mathbb{E}\left[\frac{\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right)^2}{\mathrm{Var}_{\eta}(y)^2}\right] = \frac{\mathrm{Var}\left[n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right]}{\mathrm{Var}_{\eta}(y)^2} = \frac{n\mathrm{Var}_{\eta}(y)}{\mathrm{Var}_{\eta}(y)^2} = \frac{n}{\mathrm{Var}_{\eta}(y)}. \tag{1.35}$$

Note that as n increases, the Fisher information increases linearly. The Fisher information is inversely proportional to the underlying variance. Again, by Proposition **??**, the plug-in estimator for the Fisher Information using the maximum likelihood estimator for the variance is the MLE for the Fisher information:

$$i_\eta^{(n)}(\mu)\Big|_{\eta=\hat{\eta}_{MLE}} = \frac{n}{\widehat{\text{Var}}_\eta(y)} = \mathbb{E}\left[\left(\frac{\text{d}}{\text{d}\mu}\ell(\eta)\right)^2\right] = -\mathbb{E}\left[\frac{\text{d}^2}{\text{d}\mu^2}\ell(\eta)\right]\Big|_{\eta=\hat{\eta}_{MLE}}$$

In the case of the mean parameter, we can write this as

$$= -\frac{\text{d}^2}{\text{d}\mu^2}\ell(\eta)\Big|_{\eta=\hat{\eta}_{MLE}}$$

Cramer-Rao lower bound (see Theorem **??**):

$$\text{Var}_\eta(\hat{\xi}) \geq \frac{\left(\frac{\text{d}}{\text{d}\eta}\left[\mathbb{E}_\eta(\hat{\eta})\right]\right)^2}{i_\eta^{(n)}(\xi)}$$

Any estimator that reaches (or approaches asymptotically) this lower bound is **(asymptotically) efficient**. Consider an unbiased estimator $\hat{\xi}$ ($\mathbb{E}(\hat{\xi} = \xi)$: then the bound reduces to

$$\text{Var}_\eta(\hat{\xi}) \geq \frac{1}{i_\eta^{(n)}(\xi)}$$

Further, suppose we have an unbiased estimator of the mean parameter $\hat{\mu}$. Then using (1.35) we have

$$\text{Var}_\eta(\hat{\mu}) \geq \frac{1}{i_\eta^{(n)}(\mu)} = \frac{\text{Var}_\eta(y)}{n}$$

Since we saw earlier that the MLE has exactly this variance ($\text{Var}(\overline{y}) = \text{Var}_\eta(y)/n$), it is the efficient unbiased estimator for μ. (However, it turns out that the plug-in MLE estimator for a function of this mean parameter is not in general the efficient unbiased estimator.)

4. **Exponential family model (GLM):**

   We have $Y_1, \ldots, Y_n$ i.i.d. g_η. $D_n * \eta_1, \eta_2)$

1.14.1 Regression models

Suppose $g_\eta = e^{\eta y - \psi(\eta)} \cdot c(y)$, and we observe $y_i = g_{\eta_i}$, $i \in [n]$, with $y_i \perp\!\!\!\perp y_j$ for all $i \neq j$. Then

$$g(y_1, \ldots, y_n) = \exp\left\{\sum_{i=1}^n \eta_i y_i - \psi(\eta_i)\right\} \prod_{i=1}^n c(y_i).$$

If $\psi_i = x_i'\beta \in \mathbb{R}^p$ for some β, this is a **hierarchical model**. In particular, it is an **exponential family regression model**. We must have that the natural parameter is a linear function of the covariates. If not ($\eta_i = h(x_i'\beta)$ for some nonlinear h), we get a **curved exponential family** and things are more difficult.

1.14.2 Applications—Categorical Data

1. **Poisson**

   $\eta_i = \log(\mu_i) = x_i'\beta$.

   Remark 7. Recall that for an exponential family, e.g. Poisson, if $Y \sim \text{Poisson}(\lambda)$, the natural parameter is $\eta = \log\lambda$ and the mean is λ. Then we know that $\dot{\psi}(\lambda) = \eta$ **check, it's something like that** so link function is ψ^{-1}.

   Poisson GLM: no response (logit has response) (going off p.37 of slides "m3-poisson-glm.pdf")

   choose model with high p-value (on slide 43) because significance test measures if model fit is significantly worse than saturated model. so if not true, then you want that model.

2. **Binomial** (logistic, probit). Natural parameter for binomial: $\log(\pi/(1-\pi))$. Link function:

$$\log\left(\frac{\pi}{1-\pi}\right) = x_i'\beta.$$

$$\begin{aligned}\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) &= \log\left(\frac{\mathbb{P}(Y=1 \mid X=x_i, Z=z_j)}{\mathbb{P}(Y=0 \mid X=x_i, Z=z_j)}\right) \\ &= \log\left(\frac{\mathbb{P}(Y=1, X=x_i, Z=z_j)/\mathbb{P}(X=x_i, Z=z_j)}{\mathbb{P}(Y=0, X=x_i, Z=z_j)/\mathbb{P}(X=x_i, Z=z_j)}\right) = \log\left(\frac{\mu_{i1k}}{\mu_{i0k}}\right) \\ &= (\lambda + \lambda_i^x + \lambda_1^y + \lambda_k^z + \lambda_{i1}^{xy} + \lambda_{1k}^{yz} + \lambda_{ik}^{xz}) - (\lambda + \lambda_i^x + \lambda_0^y + \lambda_k^z + \lambda_{i0}^{xy} + \lambda_{0k}^{yz} + \lambda_{ik}^{xz}) \\ &= (\lambda_1^y - \lambda_0^y) + (\lambda_{i1}^{xy} - \lambda_{i0}^{xy}) + (\lambda_{1k}^{yz} - \lambda_{0k}^{yz})\end{aligned}$$

   recall the identifiability constraints: first level is 0. So we can write this as

$$= \lambda_1^y + \lambda_{i1}^{xy} + \lambda_{1k}^{yz}$$

   If instead $\pi_i = x_i'\beta$, this is not a natural exponential family.

   Probit model:

$$\text{probit}[\pi(x)] = \Phi^{-1}(\pi(x)) = x_i'\beta \qquad \iff \pi(x) = \Phi(x_i'\beta).$$

3. **Multinomial**

1.14.3 Applications—Continuous Data

1. **Gaussian**: Link function: $\mu = \mathbb{E}(Y)$.

2. **Binomial**

3. **Multinomial**

1.15 Mixed Effects Models

y_{ij}: i cluster index, $i = 1, \dots n$, j observation number within each cluster, $j = 1, \dots, d$.

$$y_{ij} = x_i'\beta + z_{ij}v_i + \sigma\epsilon_{ij}$$

so β has to do with fixed effects. example: i is time index, j is student. so for every user you have universal characteristics for the day x_i and user-specific measurements z_{ij}. The z_{ij} effects are called **random effects**. Hierarchical modeling: suppose $z_i \in \mathbb{R}^q$ (q random effects), then we often assume $u_i \ \sim \mathcal{N}_q(0, \Sigma)$.

Can use with R `lme4` package.

1.16 Miscellaneous Topics

1.16.1 Multinomial Response

c categories in y.

$$\log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) = \text{logit}\left[\mathbb{P}\left((y_{ij} = 1 \mid y_{ij} = 1 \text{ or } y_{ic} = 1)\right] = \sum_{k=1}^{p} x_{ik}\beta_{kj} = x_i'\beta_j.\right.$$

for $j = 1, \dots, c-1$. row i: $(y_{i1}, \dots, y_{ic})$, $\sum_{k=1}^{c} y_{ik} = 1$ probabilities $(\pi_{i1}, \dots, \pi_{ic})$.

For ordinal random variables:

$$\text{logit}(\mathbb{P}(Y_i \le j)) = \log\left[\frac{\mathbb{P}(Y_i \le j)}{\mathbb{P}(Y_i > j)}\right]$$

for example, modeling size S, M, L, XL, XXL to $(1, 2, 3, 4, 5)$. Have a cumulative logit: $\mathbb{P}(Y_i \le j)$ is nondecreasing as j increases. We have

$$\text{logit}(\mathbb{P}(y_i \le j)) = \alpha_j + x_i\beta, \qquad i \in \{1, \dots, n\}, j \in \{1, \dots, c-1\}$$

under the monotonicty constraint $\alpha_1 \le \alpha_2 \le \dots \le \alpha_{c-1}$.

1.16.2 Zero-inflated response

Suppose $y_i = 0$ with probability $1 - \pi_i$ or has distribution f with probability π_i. That is, $y_i \sim (1 - \pi_i)\delta_0 + \pi_i f(; \mu_i)$. Assume $\text{logit}(\pi_i) = x_i'\beta$, and if $y_i > 0$, then $\log(\mu_i) = w_i'\gamma$ (μ_i is mean parameter). For example, f may be Poisson if y is discrete or Gaussian if y is continuous: $f = \phi(\cdot; \mu, \sigma), f(y) = \phi((y - \mu)/\sigma)$ To estimate: use maximum likelihood.

$$L(\beta;\gamma,\sigma) = \prod_{i=1}^{n}(1-\pi_i)^{I\{y_i=0\}}\left[\pi_i \cdot \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{I\{Y_i\neq 0\}} = \prod_{i=1}^{n}\left(\frac{1-\pi_i}{\pi_i}\right)^{I\{y_i=0\}}\left[\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{I\{Y_i\neq 0\}}$$

note that we can factorize this as a part that depends on β and a part that depends on γ and σ. Then log likelihood is

$$\underbrace{\sum_{i=1}^{n}\log\left(\frac{1-\pi_i}{\pi_i}\right)I\{Y_i=0\}}_{\ell(\beta)} + \underbrace{\sum_{i=1}^{n}\left(-\log(\sigma) - \frac{(y_i-\mu_i)^2}{2\sigma^2}\right)I\{Y_i\neq 0\}}_{\ell(\gamma,\sigma)}$$

and

$$\ell(\beta) = -\sum_{i:y_i=0} x_i\beta + \sum_{i=1}^{n}\log\left(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right), \qquad \ell(\gamma,\sigma) = -\sum_{i:y_i\neq 0}\left(\log(\sigma) + \frac{(y_i-\mu_i)^2}{2\sigma^2}\right)I\{Y_i\neq 0\}$$

and $\log(\mu_i) = w_i'\gamma$.

1.16.3 Overdispersion

Exponential Regression Model

$y_i \sim g_{\eta_i}$, independent. $\eta_i = x_i'\beta$, $\mu_i = \dot{\psi}(\eta_i)$. Likelihood ($L(\eta_i) = \exp(\eta_i y_i - \psi(\eta_i))\cdot c_0(y_i)$, so (ignoring last part that doesn't depend on η_i) we can write the log likelihood as $\ell(\eta_i) = \eta_i y_i - \psi(\eta_i)$.

Estimating equation: maximizing log-likelihood, the first order equations (differentiate with respect to β):

$$\frac{\partial\ell(\beta)}{\partial\beta_j} = 0, \qquad \forall j\in\{1,\dots,p\}, \qquad \frac{\partial\ell}{\partial\eta_i} = y_i - \dot{\psi}(\eta - i) = y_i - \mu_i$$

$$\sum_{i=1}^{n}\frac{\partial\ell_i(\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\ell_i}{\partial\eta_i}\cdot\frac{\partial\eta_i}{\partial\beta_j} = \sum_{i=1}^{n}(y_i-\mu_i)\cdot x_{ij} = 0, \qquad j\in\{1,\dots,p\}.$$

example use: use a quasi-likelihood approach for modeling count data where variance isn't necessarily fixed. Have over-dispersion parameter; in that case, can use different variance for count data.

Example data: Efron data set from 70s on toxoplasmosis. Reasons for overdispersion:

- subgroups in data

Remedies: GLMs in quasi likelihood ("dispersion parameter") or mixed effects/hierarchical/multi-level modeling.

Example 1.16.1 (**Correlated Binomial Trials**). $y_{ij} \sim \text{Bin}(1, \pi_i)$, $i \in [n], j \in [n_i]$. The correlation between y_{ij} and y_{ik} is ρ for $j \neq k$; $|\rho| \leq 1$. We have

$$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

1.17 Generalized linear mixed models

Nature of γ model: something like GARCH, etc. Point is, only a few parameters.

Why is MLE biased? Remember that MLE for σ^2 is biased, so basically that's why.

For REML (restricted maximum likelihood): start by projecting y onto the nullspace of X so that we don't have to worry about β:

$$y = X\beta + ZU + \epsilon,$$
$$T := (I - P_X) = (I - P_X)ZU + (I - P_X)\epsilon = (I - X(X^TX)^{-1}X^T)ZU + (I - X(X^TX)^{-1}X^T)\epsilon$$

Write down the distribution of T and the likelihood of T. Turns out log likelihood is

$$\ell_n(\gamma) = -\frac{1}{2}\left[(y - X\hat{\beta}_{GLS})^T V^{-1}(\gamma)(y - X\hat{\beta}_{GLS}) + \log\left(|V(\gamma)|\right) + \log\left(|X^TV(\gamma)X|\right)\right]$$

where again

$$\hat{\beta}_{GLS} := (X^TV(\gamma)X)^{-1}X^T(V(\gamma))^{-1}y.$$

(comes up automatically as a result of projection; not here because of an iterative procedure).

Testing:

1. **Fixed effects** (β): Partition $X = [X_0|X_1]$, $\beta = (\beta_0, \beta_1)$. $H_0 : \beta_1 = 0$. Likelihood ratio test:

$$2\left[\underbrace{\ell(\hat{\beta}, \hat{\sigma}^2; \hat{V})}_{\text{complete model}} - \underbrace{\ell(\hat{\beta}_0, \hat{\sigma}_0^2; \hat{V}_0)}_{\text{null model}}\right] \xrightarrow{d} \chi^2_{\dim(\beta_1)}.$$

   Issues: test can be anti-conservative (probability of type 1 error higher than stated α); asymptotics "kick in" very late. Parametric bootstrap can work better.

2. **Random effects** (U): $H_0 : \sigma_j = 0$ (mean of random effects is 0 by assumption, so if variance is also 0, effect is "not there"). Also do likelihood ratio test, in the end turns out to also be asymptotically χ^2.

Composite effect: both fixed and random effects. (some economic models only allow random effects if there is also a fixed effect—enforce hierarchy).

1.17.1 Longitudinal data analysis and Generalized Estimating Equations

t for time, i for user/customer. y_{it} has correlation across time. $\mathbb{E}y_{it} = \mu_{it}$, $\text{Var}(y_{it}) = \phi a(\mu_{it})$, $g(\mu_{it}) = X_{it}'\beta + z_{it}'\gamma$ where a can be any arbitrary function (known variance function). If $\phi = 1$, no overdispersion. Estimate ϕ via quasi-likelihood. This case has no correlation across y_{it}. One way to introduce correlation (and parameterize a):

$$\text{Var}(y_{it}) = \phi A_i^{1/2} R(\alpha) A_i^{1/2}$$

where $A_i = \text{diag}(a(\mu_i t))$. For example, for a moving average $MA(1)$ process, we have

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \alpha & 1 & \alpha & \cdots & 0 \\ 0 & \alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

1.18 Causal Inference

1.18.1 Factorial Design (see R lab 7)

2 or more discrete factors. Fully cross-classified.

$$y_{ijk} = \mu + X\beta + \underbrace{\alpha_i}_{\text{school}} + \underbrace{\beta_j}_{\text{class}} + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Nested design:

$$y_{ijk} = \mu + X\beta + \alpha_i + \beta_{ij} + \epsilon_{ijk}.$$

Imbalanced design:

$$y_{ijk} = \underbrace{\mu}_{\text{general effect}} + \underbrace{\alpha_i}_{\text{treatement effect}} + \underbrace{\beta_j}_{\text{block effect}} + \epsilon_{ijk}$$

Complete vs. incomplete design: all treatments do not occur in the same block. Treatment contrast: difference between treatment effect i and j: $\alpha_i - \alpha_j$. BIBD: balanced incomplete block designs. All treatment contrasts are estimable. I treatments, J blocks. Each treatment occurs exactly r times in the design. $n = rI = kJ$. Typically $r \ll J$.

1.19 Math 547

Exercise 3.16: this inequality talks about the number of misclassifications, not the probability of misclassification under any distribution.

1.19.1 Perceptron Algorithm

Remark 8. Note that the run times only depends on the ℓ_2 norm of the solution loadings w and the ℓ_2 norm of the longest vector in the data set. That sounds good since it doesn't depend on the size of the data, but in the worse case θ can grow exponentially in the dimension of the data.

Also, note that the actual run time is at least linear in the size of the data, since on each iteration the algorithm checks some multiple of m points.

1.19.2 Mercer's Theorem

How is this an infinite-dimensional version of the Exercise? Let M be a $k \times k$ real symmetric matrix. By the Spectral Theorem, there exists an orthogonal Q and a diagonal D such that $M = Q^T D Q$. For all $1 \leq p \leq k$, let λ_p denote the pth diagonal entry of D. Let $\psi_i^p \in \mathbb{R}^k$ denote the ith row of Q. Then

$$m_{ij} = \sum_{p=1}^{k} \lambda_p \psi_i^{(p)} \psi_j^{(p)}, \qquad \forall 1 \leq i, j \leq k.$$

Also, $m(x, y)$ is called a **kernel.**

How to get ψ from $m(x, y)$? Definte

$$\ell_2 := \{\}$$

Note: if we could write the algorithm in terms of $m(x, y)$, we don't need to specify this embedding ϕ at all.

1.20 Norms

Proposition 1.20.0.1. Let $C \subseteq \mathbb{R}^n$ be a symmetric (if $c \in C$ then $-c \in C$) convex set containing 0. Define for all $x \in \mathbb{R}^n$,

$$\|x\|_C := \inf \left\{ t > 0 : \frac{x}{t} \in C \right\}$$

Then $\|x\|_C$ is a norm.

Theorem 1.20.0.2 (**notes from proof of lemma 6.24**).

$$d_{2,\mathbb{P}}(f,g)^2 - d_{2,\mathbb{P}_m}(f,g)^2 = (\mathbb{E}_{\mathbb{P}}|f-g|^2)^{1/2} - (\mathbb{E}_{\mathbb{P}_m}|f-g|^2)^{1/2} = \mathbb{E}h(X) - \frac{1}{m}\sum_{i=1}^{m} h(X_i) = \frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}h(X) - h(X_i))$$

note from proof of theorem 6.23: needed to bound number of points in order to apply Sauer-Shelah lemma (only applies for finite number of points)

Definition 1.20.1 (**Nuclear norm**). Let $A \in \mathbb{R}^{m_1 \times m_2}$. The **nuclear norm** of A, denoted $\|A\|_*$ is given by

$$\|A\|_* := \sum_{j=1}^{\text{rank}(A)} \sigma_j(A),$$

where $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{\text{rank}(A)}(A)$ are the singular values of A.

Definition 1.20.2 (**Spectral norm** or **operator norm**). Let $A \in \mathbb{R}^{m_1 \times m_2}$. The **Spectral norm** or **operator norm** of A is given by

$$\|A\| := \max_j \sigma_j(A).$$

Exercise 1. Prove that the operator norm is the dual of the nuclear norm. (For any norm $\|\cdot\|'$, its dual norm is defined as

$$\|v\|'_* := \sup_{u:\|u\|'\leq 1} \langle u, v\rangle.)$$

That is, prove

$$\|A\| = \sup_{\|B\|_*\leq 1} \text{Tr}(A^T B).$$

1.21 Collaborative Filtering and Trace Regression (Math 541B)

Notation (Netflix problem): m_1 movies, m_2 users, matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$ (rows correspond to movies, columns correspond to users), $\{A_0\}_{ij} \in [5]$.

1.21.1 Trace Regression

Let $A_1, A_2 \in \mathbb{R}^{m_1 \times m_2}$. Define $\langle A_1, A_2\rangle := \text{Tr}(A_1^T A_2) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(A_1)_{ij}(A_2)_{ij}$ (equivalent to vectorizing each matrix, then taking inner product). Assume that (X, Y) is a random sample with $X \in \mathbb{R}^{m_1 \times m_2}, Y \in \mathbb{R}$ such that $Y = \langle A_0, X\rangle + \xi$ where $A_0 \in \mathbb{R}^{m_1 \times m_2}$, $X \in \mathbb{R}^{m_1 \times m_2}$ is a random measurement matrix, and $\xi \in \mathbb{R}$ is random noise with $\mathbb{E}\xi = 0$, independent of X by assumption.

Example 1.21.1. Suppose A_0 and X are diagonal; that is, $A_0 = \text{diag}(a_0)$, $X = \text{diag}(x)$, then $\langle A_0, X\rangle = \langle a_0, x\rangle = \sum_{j=1}^{m_1}(a_0)_j x_j$. (This is a standard linear model.)

Example 1.21.2 (**Low-rank matrix completion**)**.** Suppose $A_0 \in \mathbb{R}^{m_1 \times m_2}$ has low rank and X has uniform distribution on the set

$$\mathcal{E} := \{e_i(m_1)e_j(m_2)^T, i \in [m_1], j \in [m_2]\}$$

where $e_i(m)$ has the same dimension as m and contains 1 in the ith entry and 0 in every other entry. That is, $e_i(m_1)e_j(m_2)^T$ is a matrix containing a 1 only in entry (i, j) and 0 elsewhere, and

$$\langle A_0, e_i(m_1)e_j(m_2)^T \rangle = (A_0)_{ij}.$$

(Note that it is important that

$$\mathbb{P}(X = e_i e_j^T) > 0 \qquad \forall i, j.$$

To see why, assume that we only observe movie ratings from one user. Even if that user rates every single movie and A_0 has rank 1, we still will not be able to make any reasonable predictions about the other users' ratings.) So our model is

$$Y_j = \langle A_0, X_j \rangle + \xi_j, \qquad X_j \sim \text{Uniform}(\mathcal{E}), \qquad \xi_j \perp\!\!\!\perp X_j, \qquad j \in [n].$$

By Exercise 1,

$$\langle A_1, A_2 \rangle \leq \|A_1\| \cdot \|A_2\|_*.$$

(This fact is sometimes known as *trace duality.*) The least squares estimator of A_0 is

$$\hat{A} := \underset{S \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2.$$

But since $n \ll m_1 m_2$, the solution is not unique; this has too many degrees of freedom. (One solution is

$$\hat{A} = \frac{1}{n} \sum_{j=1}^{n} Y_j X_j,$$

but this is not low-rank.) (Exercise: calculate $\mathbb{E}\|\hat{A} - A_0\|_F^2$.) One idea is to impose a constraint that $\text{rank}(S) \leq k$. Then the problem becomes

$$\begin{aligned} \hat{A} = \underset{S \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \quad & \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2 \\ \text{subject to} \quad & \text{rank}(S) \leq k, \qquad k \geq \text{rank}(A_0) \end{aligned}$$

We know that $A_0 \in \mathcal{A} = \{A : \text{rank}(A) \leq k, \|A\|_* \leq \|A_0\|_*\}$. But this problem is non-convex. The idea is to replace $\mathcal{A}$ by its convex hull. Consider the set $\tilde{\mathcal{A}} \subset \mathcal{A}$ defined by

$$\tilde{\mathcal{A}} = \left\{\pm\|A_0\|_* = uv^T, u \in \mathbb{R}^{m_1}, v \in \mathbb{R}^{m_2}, \|u\|_2 = 1, \|v\|_2 = 1\right\}.$$

We claim that the convex hull of $\tilde{\mathcal{A}}$ is equal to

$$\|A_0\|_* \cdot B(0, 1, \|\cdot\|_*)$$

where

$$B(0, 1, \|\cdot\|_*) = \{S \in \mathbb{R}^{m_1 \times m_2} : \|S\|_* \leq 1\}.$$

and is also the convex hull of $/\mathcal{A}$. Since the set $\|A_0\|_* \cdot B(0, 1, \|\cdot\|_*)$ contains A and is convex, the convex hull of $\mathcal{A}$ is equal to $\|A_0\|_* \cdot B(0, 1, \|\cdot\|_*$. (We took $\tilde{\mathcal{A}} \subset \mathcal{A}$ and showed that $\mathcal{A} \subset \text{co}(\tilde{\mathcal{A}})$. Since $\tilde{\mathcal{A}} \subset \mathcal{A}$, $\text{co}(\tilde{\mathcal{A}}) \subseteq \text{co}(\mathcal{A})$, so we must have $\text{co}(\tilde{\mathcal{A}}) = \text{co}(\mathcal{A})$.) So we can change our optimization problem to the convex problem

$$\begin{aligned} \hat{A} = \underset{S \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \quad & \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2 \\ \text{subject to} \quad & \|S\|_* \leq \|A_0\|_*. \end{aligned}$$

Lastly, since the nuclear norm of A_0 is unknown, we will replace $\|A_0\|_*$ with a sufficiently large constant:

$$\begin{aligned} \hat{A} = \underset{S \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \quad & \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2 \\ \text{subject to} \quad & \|S\|_* \leq t \end{aligned}$$

for some $t > 0$. In practice, we can look at the Lagrangian form:

$$\hat{A}_\lambda := \underset{S \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2 + \lambda \|S\|_* \right\}$$

for some $\lambda \in \mathbb{R}_+$. (Nuclear norm penalization is a convex relaxation.) Consider:

$$\frac{1}{n} \sum_{j=1}^{n} (Y_j - \langle S, x_j \rangle)^2 + \lambda \|S\|_* = \frac{1}{n} \sum_{j=1}^{n} Y_j^2 + \frac{1}{n} \sum_{j=1}^{n} \langle S, x_j \rangle^2 - 2 \left\langle \frac{1}{n} \sum_{i=1}^{n} X_j Y_j, A \right\rangle + \lambda \|S\|_*$$

Consider the term

$$+ \frac{1}{n} \sum_{j=1}^{n} \langle S, x_j \rangle^2.$$

By the Law of Large Numbers, this converges to its expectation, so we can replace it with its expectation

$$\mathbb{E}\langle S, x_j\rangle^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{1}{m_1 m_2}\underbrace{\langle A_i e_i(m_1) e_j(m_2)\rangle^2}_{A_{ij}} = \frac{1}{m_1 m_2}\sum_{i,j}A_{ij}^2 = \frac{1}{m_1 m_2}\|A\|_F^2.$$

The problem becomes

$$\begin{aligned}\hat{A} = \arg\min_A \left\{\frac{1}{m_1 m_2}\|A\|_F^2 - 2\left\langle \frac{1}{n}\sum_{i=1}^n X_j Y_j, A\right\rangle + \lambda\|S\|_*\right\} \\ = \arg\min_A \left\{\|A\|_F^2 - 2\left\langle \frac{m_1 m_2}{n}\sum_{i=1}^n X_j Y_j, A\right\rangle + \lambda'\|S\|_*\right\} \\ = \arg\min_A \left\{\|\tilde{X} - A\|_F^2 + \lambda'\|S\|_*\right\}\end{aligned}$$

where

$$\tilde{X} := \frac{m_1 m_2}{n}\sum_{j=1}^n Y_j X_j.$$

Observe that $\tilde{X}$ is an unbiased estimator of A:

$$\begin{aligned}\mathbb{E}\tilde{X} = \frac{m_1 m_2}{n} n\mathbb{E}YX = m_2 m_2 \mathbb{E}(\langle A_0, X\rangle X + \xi X) = m_2 m_2 \mathbb{E}(\langle A_0, X\rangle X) \\ = m_2 m_2 \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(A_0)_{ij} e_i(m_1) e_j(m_2) = A_0\end{aligned}$$

since $\xi \perp\!\!\!\perp X$ and $\mathbb{E}\xi = 0$. (So if the penalty is 0, we have an unbiased estimator for A_0, but we want the penalty because the unbiased estimator is bad.)

Theorem 1.21.1.1. $\hat{A}_\lambda$ is given explicitly by the following:

$$\hat{A}_\lambda = \sum_{j=1}^{\min(m_1,m_2)}\left(\sigma_j(\tilde{X}) - \frac{\lambda m_1 m_2}{2}\right)_+ u_i(\tilde{X}) v_j(\tilde{X})$$

where $x_+ := \max\{0, x\}$ is the soft-thresholding function. (summary: take SVD of $\tilde{X}$, then apply soft-thresholding to the singular values. soft singular value thresholding.)

Proof. Consider

$$\hat{a}_\lambda = \arg\min_{a\in\mathbb{R}}\left[(a-x)^2 + \lambda|a|\right].$$

Then

$$0 \in \partial F(\hat{a}_\lambda) \iff \exists v \in \partial|\hat{a}_\lambda| : 2(\hat{a}_\lambda - x) + \lambda v = 0$$

$$\partial F(a) = \{2(a-x) + \lambda v, v \in \partial|a|\} \qquad \implies \hat{a}_\lambda = x - \frac{\lambda}{2}v$$

Two possibilities:

(a)
$$|x| > \frac{\lambda}{2} :$$
take $v = \text{sign}(x)$. Then $\hat{a}_\lambda = (|x| - \lambda/2)\,\text{sign}(x)$.

(b) $|x| \leq \lambda/2$: If $\hat{a}_\lambda > 0, v = 1$. $\hat{a}_\lambda = x - \lambda/2 < 0$. If $\hat{a}_\lambda < 0, v = -1$. Then $\hat{a}_\lambda = x + \lambda/2 > 0$: contradiction. So $\hat{a}_\lambda = 0$.

$\hat{a}_\lambda = (|x| - \lambda/2)_+ \,\text{sign}(x)$.

□

Fact: let $\|\cdot\|'$ be any norm. Then

$$\partial\|x\|' = \begin{cases} y : \|y\|'_* = 1, \langle y, x\rangle = \|x\|', & x \neq 0, \\ y : \|y\|'_* \leq 1, & x = 0, \end{cases}$$

where $\|\cdot\|'_*$ is the dual norm of $\|\cdot\|'$. Apply this fact to the nuclear norm to get that

$$\partial\|A\|_* = \left\{ \sum_{j=1}^{\text{rank}(A)} u_j(A)v_j(A)^T + \sum_{j=\text{rank}(A)}^{\min(m_1,m_2)} w_1 u_j(A)v_j(A)^T, \qquad w_j \in [-1,1] \right\}.$$

(since singular values are always non-negative, so subdifferntial is 1 when singular value is positive and any element of $[-1,1]$ when singular value is 0.) Let

$$F(A) := \|A - \tilde{X}\|_F^2 + \lambda m_1 m_2 \|A\|_*.$$

A_λ minimizes $F(A)$ if and only if $0 \in \partial F(\hat{A}_\tau)$. Note that

$$\partial F(\hat{A}_\lambda) = \left\{ 2(\hat{A}_\lambda - \tilde{X}) + \lambda m_1 m_2 v, \qquad v \in \partial\|\hat{A}_\lambda\|_* \right\}.$$

(a)
$$\sigma_j(\tilde{x}) > \frac{\lambda}{2} m_1 m_2 \implies \sigma_j(\hat{A}_\lambda) = \sigma_j(\tilde{x}) - \frac{\lambda}{2} m_1 m_2.$$

(b)

$$\sigma_j \leq \frac{\lambda}{2} m_2 m_2 \implies \sigma_j(\hat{A}_\lambda) = 0.$$

Question: which choice of w_j's are required here?

Theorem 1.21.1.2. Assume that $(X_j, Y_j), j \in [n]$ are i.i.d., $X_j \sim \text{Uniform}(e)$, $|Y_j| \leq \eta$ (maximal rating) with probability 1. Let

$$M = \frac{1}{n} \sum_{j=1}^{n} \left(Y_j X_j - \underbrace{\frac{A_0}{m_1 m_2}}_{\mathbb{E}(Y_j X_j)} \right).$$

Assume that $\lambda \geq 2\|M\|_1$ (twice the largest singular value of the matrix), and $n \gg \min\{m_1, m_2\} \log(m_1 + m_2)$ (tells us the number of ratings we have observed exceeds at least the minimal dimension of the matrix by some logarithmic factor; have observed a rating for each movie and each user at least once). Then the following holds:

$$\|\hat{A}_\lambda - A_0\|_F^2 \leq \left(\frac{1+\sqrt{2}}{2} \right)^2 m_1 m_2 \lambda^2 \operatorname{rank}(A_0).$$

(as long as λ is big enough, then estimator performs sufficiently well.)

Proposition 1.21.1.3. Let

$$\lambda \geq 4\eta \sqrt{\frac{t + \log(m_1 + m_2)}{\min(m_1, m_2) n}}.$$

Then $\lambda \geq 2\|M\|$ with probability greater than or equal to $1 - e^{-t}$.

With these two results, we get that for

$$\lambda = 4\eta \sqrt{\frac{t + \log(m_1 + m_2)}{\min(m_1, m_2) n}},$$

we get

$$\|\hat{A}_\lambda - A_0\|^2 \leq \underbrace{\left(\frac{1+\sqrt{2}}{2} \right)^2 16}_{\mathbb{C}} \eta^2 m_1 m_2 \cdot \frac{t + \log(m_1 + m_2)}{\min(m_1, m_2) n} \operatorname{rank}(A_0)$$

$$= \frac{\max\{m_1, m_2\} \operatorname{rank}(A_0)}{n} \mathbb{C} \eta^2 (t + \log(m))$$

achieves best bound without knowing rank of matrix in advance.

1.22 Dynamic Programming

1.22.1 Introduction to Dynamic Programming and Principle of Optimality (Sections 1.1 - 1.3 of [Bertsekas, 2012a])

Discrete-time dynamic system:

1. **System dynamics:**
   - x_k: state at time k.
   - u_k: control/decision in period k. Policy: mapping between state and action you will take (a policy is a state-dependent decision, not a decision).
   - w_k: random noise. May depend on x_k, u_k, k itself. But **conditioned on** x_k, U_k, we assume w_k is independent of $w_{k-1}, \dots, w_1$.

   Model: $x_{k+1} = f_k(x_k, u_k, w_k)$.

2. **Additive Cost Structure:**

   For period $k = 0, 1, \dots, K-1$, cost $g_k(x_k, u_k, w_k)$ in period k. Then terminal cost $g_N(x_N)$. Goal: choose a policy to minimize total expected cost

$$\mathop{\mathbb{E}}_{w_0,\dots,w_{N-1}}\left[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) + g_n(x_n)\right].$$

Example 1.22.1 (**Inventory control (Example 1.1.1 in [Bertsekas, 2012a] , p.5)l**). 1. Suppose we have a horizon of N periods. In each period k, x_k = number of units of the product available to sell. (We allow $x_k \leq 0$.) $u_k \geq 0$ is the amount ordered from the supplier in period k. (Assume instantaneous delivery, so zero lead time.) Finally, w_k would be the demand for the product in period k. We will assume $w_k \perp\!\!\!\perp w_j$ for all $k \neq j$. Note that $x_{k+1} = x_k - w_k + u_k =: f_k(x_k, w_k, u_k)$.

2. Costs:
   - ordering cost (say $C \cdot u_k$).
   - inventory cost: i per unit per period.
   - Backorder cost: b per unit per period.

   So

$$g_k(x_k, u_k, w_k) = cu_k + i \cdot \underbrace{x_k^+}_{\max\{x_k, 0\}} + b \cdot \underbrace{x_k^-}_{\max\{-x_k, 0\}} = cu_k + i \cdot (x_k + u_k - w_k)^+ + b(w_k - x_k - u_k)^+$$
$$= cu_k + i(x_{k+1})^+ + b(-x_{k+1})^+.$$

   Different notation:

$$\mathcal{G}(x_k, u_k, w_k) = \underbrace{p}_{\text{backorder cost/period}} \cdot \max\{0, -x\} + \underbrace{i}_{\text{inventory cost/period}} \cdot \max\{x, 0\},$$

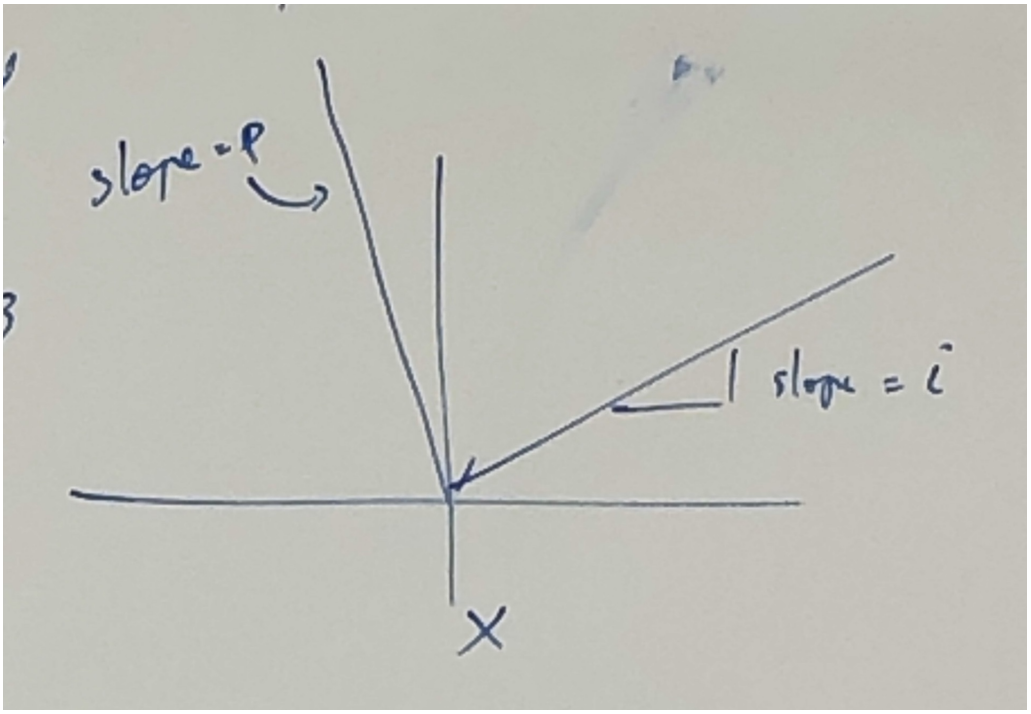

Figure 1.5: Depiction of cost structure for Example 1.22.1. Note that this function $x \mapsto \mathcal{G}(x)$ is convex in x; this is relevant in the proof of Theorem 1.22.3.1.

ordering cost c/unit, assume $p > c$ (see Figure 1.5).

Objective: Minimize

$$\mathop{\mathbb{E}}_{w_0,\ldots,w_{N-1}}\left[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) + g_n(x_n)\right].$$

or

$$\mathop{\mathbb{E}}_{w_0,\ldots,w_{N-1}}\left[\sum_{k=0}^{N-1} c \cdot u_k + \mathcal{G}(x_k + u_k - w_k)\right] = \mathop{\mathbb{E}}_{w_0,\ldots,w_{N-1}}\left[\sum_{k=0}^{N-1} c \cdot u_k + p(w_k - x_k - u_k)^+ + i(x_k + u_k - w_k)^+\right].$$

DP Equation: $J_k(x_k)$: minimum (over all possible policies) expected cost-to-go given that the state at the beginning of the period is x_k:

$$J_k(x_k) = \min_{\Pi^k = (\pi_k, \ldots, \pi_{N-1})} \left\{ \mathop{\mathbb{E}}_{w_0,\ldots,w_{N-1}}\left[\sum_{\ell=k}^{N-1} c \cdot \pi_\ell(x_\ell) + \mathcal{G}(x_\ell + \pi_\ell(x_\ell) - w_\ell)\right]\right\}.$$

So, for $k \in N-1, N-2, \ldots, 0$,

$$J_k(x_k) = \min_{u_k \geq 0} \left\{ c \cdot u_k + \mathop{\mathbb{E}}_{w_k} \left[\mathcal{G}(x_k + u_k - w_k) \right] + \mathop{\mathbb{E}}_{w_k} \left[J_{k+1}(x_k + u_k - w_k) \right] \right\}.$$

Open loop policy: Determine $u_0, u_1, \ldots, u_{N-1}$ in advance.

Definition 1.22.1. Closed loop policy: u_k is a function of the state x_k.

In particular, let $\mu_k : \mathbb{R} \to \mathbb{R}_+$ be our policy. Then $\mu_k(x_k)$ is the order quantity in period k given that x_k units remain.

Definition 1.22.2 (**Policy**)**.** A **policy** is a sequence of functions

$$\pi = \{\mu_0(\cdot), \ldots, \mu_{N-1}(\cdot)\}$$

mapping the state at state k to an action. The cost associated with the policy π given an initial state x_0 is denoted by

$$J_\pi(x_0) = \mathbb{E}\left[\sum_{k=0}^{N-1} \left(c \cdot \mu_k(x_k) + i \cdot (x_k + \mu_k(x_k) - w_k)^+ + b(w_k - x_k - \mu_k(x_k))^+ \right) + g_n(x_n) \right].$$

Our goal is to find an optimal policy

$$\pi^* \in \mathop{\arg\min}_{\pi \in \Pi} \{J_\pi(x_0)\}.$$

Ideally we would like this policy to be uniformly optimal over all possible x_0. (Next week: we'll show that an order-up-to policy is optimal for this problem; that is, for each period k there exists a constant $S_k \geq 0$ such that $\mu_k^*(x_k) = [S_k - x_k]^+$. That is, if the amount of inventory you have is less than S_k, order up to that number; otherwise, don't order.)

Basic ingredients:

(a) Discrete-time system: $x_{k+1} = f_k(x_k, u_k, w_k)$, $k = 0, 1, \ldots, N-1$.

(b) Independent random noise w_k: w_k can depend on x_k, u_k.

(c) Constraints on control: $u_k \in \mathcal{U}_k(x_k)$.

(d) Additive cost.

(e) Closed loop policy.

Note: when the number of states in the system is discrete/finite, instead of writing $x_{k+1} = f_k(x_k, u_k, w_k)$, you can instead of writing this formula specify transition probabilities, e.g.

$$\mathbb{P}\left(x_{k+1} = j \mid x_k = i, u_k = u\right) = P_{ij}(u, k)$$

(i.e. this suffices; all you need is the transition probabilities.)

Note that $J^* : \mathcal{X} \to \mathbb{R}$ is called an **optimal value function.** We would like an algorithm to find J^* (and π^*). Dynamic programming algorithm: the Bellman equations (or DP equations) relies on the principle of optimality.

Definition 1.22.3 (**Principle of optimality (Section 1.3, p. 20 of [Bertsekas, 2012a])**). Suppose $\pi^* = \{\mu_0^*(\cdot), \dots, \mu_{N-1}^*(\cdot)\}$ is an optimal policy. Suppose π^* is used, and a given state x_i occurs in point i with a positive probability. Consider a sub-problem where we are at x_i in period i and wish to minimize the cost-to-go from time i to the end of the horizon (time N),

$$\mathbb{E}\left[\sum_{k=i}^{N-1} \left(c \cdot \mu_k(x_k) + i \cdot (x_k + \mu_k(x_k) - w_k)^+ + b(w_k - x_k - \mu_k(x_k))^+\right) + g_n(x_n)\right].$$

Then the truncated policy $\{\mu_i^*(\cdot), \dots, \mu_{N-1}^*(\cdot)\}$ is also optimal for this sub-problem.

Example 1.22.2 (Analogy)**.** Consider the shortest path from LA to New York. If that path passes through Chicago, then the shortest path from Chicago to New York must be identical to that part of the shortest path from LA to New York.

Implications of the principle of optimality: start with a sub-problem of length 0: use the one at the beginning of period N. Then the cost-to-go is $J_N(x_N) = g_N(x_N)$.

Consider a tail sub-problem of length 1. Suppose we are in state x_{N-1}. Then the optimal policy going forward is

$$\arg\min_{u_{N-1} \geq 0} \left\{c \cdot u_{N-1} + \mathbb{E}\left[i \cdot (x_{N-1} + u_{N-1} - w_{N-1})^+ + b(w_{N-1} - x_{N-1} - u_{N-1})^+ + J_N(x_N)\right]\right\}$$

We can continue backwards:

$$J_{N-2}(x_{N-2}) = \min_{u_{N-1} \geq 0} \left\{c \cdot u_{N-2} + \mathbb{E}\left[i \cdot (x_{N-2} + u_{N-2} - w_{N-2})^+ + b(w_{N-2} - x_{N-2} - u_{N-2})^+ + J_{N-1}(x_{N-1})\right]\right\}$$

$$J_k(x_k) = \min_{u_{k+1} \geq 0} \left\{c \cdot u_k + \mathbb{E}\left[i \cdot (x_k + u_k - w_k)^+ + b(w_k - x_k - u_k)^+ + J_{k+1}(x_{k+1})\right]\right\}$$

and work our way back to $J_0(x_0)$.

Sketch of proof: consider the following backward sequence of functions: $J_N(x_N) = g_N(x_N)$. For $k = N - 1, N - 2, \dots, 2, 1, 0$,

$$J_k(x_k) = \min_{u_k \in \mathcal{U}_k(x_k)} \left\{\mathbb{E}\left[g_k(x_k, u_k, w_k) + J_{k+1}\left(f_k(x_k, u_k, w_k)\right)\right]\right\} \tag{1.36}$$

Theorem 1.22.1.1 (**Optimality of Dynamic Programming Algorithm (Proposition 1.3.1 in [Bertsekas, 2012a], p. 25)**). For every initial state x_0, the optimal cost $J^*(x_0)$ is equal to $J_0(x_0)$. Moreover, if $\mu_k^* = \mu_k^*(x_k)$ minimizes the right hand side of (1.36), then $\pi^* = \{\mu_0^*(\cdot), \dots, \mu_{N-1}^*(\cdot)\}$ is an optimal policy.

Note: μ_k^* is a function that maps the state at time k to an action. u_k is that action.

Proof. For any admissible policy $\pi = \{\mu_0(\cdot), \dots, \mu_{N-1}(\cdot)\}$, let $\pi^k = \{\mu_k(\cdot), \dots, \mu_{N-1}(\cdot)\}$ for the tail subproblem of length $N-k$ (starting from period k). For $k = 0, 1, \dots, N-1$, let $J_k^*(x_k)$ be the minimum cost-to-go for the $N-k$ stage problem given that we are in state x_k at time k. That is,

$$J_k^*(x_k) = \min_{\pi^k = \{\mu_k(\cdot), \dots, \mu_{N-1}(\cdot)\}} \left\{ \mathbb{E}\left[\sum_{i=k}^{N-1} g_i(x_i, \mu(x_i), w_i) + g_N(x_N) \right] \right\}$$

where $g_i(x_i, \mu(x_i), w_i) = (c \cdot \mu_i(x_i) + i \cdot (x_i + \mu_i(x_i) - w_i)^+ + b(w_i - x_i - \mu_i(x_i))^+)$. (Note: $J_0^*(x_0) = J^*(x_0)$. We'll prove by induction that $J_k^*(x_k) = J_k(x_k)$ for all $k = N, N-1, \dots, 2, 1, 0$.

Base case: is it true that $J_N^*(x_N) = J_N(x_N)$? Yes; just have to minimize $g_N(x_N)$.

Inductive step: Suppose $J_{k+1}^*(x_{k+1}) = J_{k+1}(x_{k+1})$ for all x_{k+1}. We'll now show that $J_k^*(x_k) = J_k(x_k)$ for all x_k. Note that

$$\begin{aligned}
J_k^*(x_k) &= \min_{(\mu_k(\cdot), \pi^{k+1})} \left\{ \mathbb{E}\left[g_k(x_k, \mu(x_k), w_k) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu(x_i), w_i) + g_N(x_N) \right] \right\} \\
&= \min_{\mu_k(\cdot)} \left\{ \mathbb{E}\left[g_k(x_k, \mu(x_k), w_k) \right] + \min_{\pi^{k+1}} \left\{ \mathbb{E}\left[\sum_{i=k+1}^{N-1} g_i(x_i, \mu(x_i), w_i) + g_N(x_N) \right] \right\} \right\} \\
&= (\text{by definition}) \min_{\mu_k(\cdot)} \left\{ \mathbb{E}\left[g_k(x_k, \mu(x_k), w_k) \right] + J_{k+1}^*(x_{k+1}) \right\} \\
&= (\text{by inductive hypothesis}) \min_{\mu_k(\cdot)} \left\{ \mathbb{E}\left[g_k(x_k, \mu(x_k), w_k) \right] + J_{k+1}(x_{k+1}) \right\} \\
&= \min_{u_k \in \mathcal{U}_k(x_k)} \left\{ \mathbb{E}\left[g_k(x_k, \mu(x_k), w_k) \right] + J_{k+1}(f_k(x_k, \mu_k(x_k), w_k)) \right\},
\end{aligned}$$

verifying that (1.36) is the minimal cost.

□

1.22.2 State Augmentation and Other Reformulations (Section 1.4 of [Bertsekas, 2012a])

- **Time lag:** Suppose that $x_{k=1}$ depends on x_k, u_k, and x_{k-1}, w_{k-1}. Then $x_{k+1} = f(x_k, u_k, w_k, x_{k-1}, w_{k-1})$. Trick: augment the state variable: $\tilde{x_k} = (x_k, x_{k-1}, w_{k-1})$. Then $\tilde{x}_{k+1} = \tilde{f}(\tilde{x}_k, u_k, w_k)$; in particular,

$$\begin{pmatrix} x_{k+1} \\ x_k \\ w_{k-1} \end{pmatrix} = \begin{pmatrix} f(x_k, u_k, w_k, x_{k-1}, w_{k-1}) \\ x_k \\ w_{k-1} \end{pmatrix}$$

- **Forecast:** Suppose at time k we observe a forecast y_k that influences our assessment of the probability distribution of w_k. In particular, suppose w_k can have one of m possible distributions $Q_1(\cdot), \dots, Q_m(\cdot)$. Then the forecast is $y_k \in \{1, \dots, m\}$. Assume y_k is exogenous. Suppose we have a random variable

ξ_k such that if $\xi_k = i$, then w_{k+1} occurs according to probability distribution Q_i, and $\mathbb{P}(\xi_k = i) = p_i$. The state at time k is then $\tilde{x}_k = (x_k, y_k)$, and

$$\tilde{x}_{k+1} = \begin{pmatrix} x_k \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} f_k(x_k, u_k, w_k) \\ \xi_k \end{pmatrix} = \tilde{f}_k((x_k, y_k); u_k, w_k).$$

Then

$$J_k(x_k, y_k) = \min_{u_k \in \mathcal{U}_k} \mathop{\mathbb{E}}_{w_i \sim Q(\cdot)} \left[g_k(x_k, y_k, w_k) + \underbrace{\sum_{j=1}^{m} \mathbb{P}(\xi_{k+1} = j) \cdot J_{k+1}(x_{k+1}, j)}_{\text{expected cost-to-go}} \right].$$

- **Correlated disturbances:** $w_k = \lambda w_{k-1} + \xi_k$, where $\xi_0, \dots, \xi_{N-1}$ are i.i.d. Then (letting $y_k = w_{k-1}$)

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} f_k((x_k, w_k); u_k, \lambda w_{k-1} + \xi - k) \\ \lambda y_k + \xi_k \end{pmatrix}$$

- **Removal of uncontrollable states:** Suppose the state is described by (x_k, y_k), where $x_{k+1} = f_k((x_k, y_k), u_k, w_k)$ and the evolution of y_k is generated by $\mathbb{P}_k(y_k \mid x_k)$ and is independent of the control (except from the state x_k). Standard dynamic program:

$$J_k(x - k, y - k) = \min_{u \in \mathcal{U}_k} \mathbb{E}\left[g_k(x_k, y_k, u_k, w_k) + \mathbb{E}\left(f_k(x_k, y_k, u_k, w_k) \right) \right]$$

We would like to reduce the dimension of the state space since y_k is dependent on x_k (and a probability distribution). One way to do this is as follows:

$$\hat{J}_k(x_k) = \mathbb{E}_{y_k}\left[J_k(x_{k+1}, y_{k+1}) \mid x_k \right] = \sum_a \mathbb{P}(y_{k+1} = a \mid x_k) \cdot J_k(x_k, a).$$

Question: how is $\hat{J}_k(\cdot)$ related to $\hat{J}_{k+1}(\cdot)$?

$$\begin{aligned}
\hat{J}_h(x_h) &= \mathbb{E}_{y_h}\left[J_h(x_h, y_h) \mid x_h \right] \\
&= \mathbb{E}_{y_h}\left[\min_{u_k} \left\{ \mathbb{E}_{w_h, x_{k=1}, y_{k+1}} \left(g_k(x_k, y - k, u_k, w_k) + J_{k+1}(x_{k+1}, y_{k+1}) \mid x_k, y_k, u_k \right) \right\} \mid x_h \right] \\
&= \mathbb{E}_{y_h}\left[\min_{u_k} \left\{ \mathbb{E}_{w_h, x_{k=1}} \left(g_k(x_k, y - k, u_k, w_k) + \mathbb{E}\left[J_{k+1}(x_{k+1}, y_{k+1}) \mid x_k, y_k, w_k \right] \mid x_k, y_k, u_k \right) \right\} \mid x_h \right] \\
&= \mathbb{E}_{y_h}\left[\min_{u_k} \left\{ \mathbb{E}_{w_h, x_{k=1}} \left(g_k(x_k, y - k, u_k, w_k) + \hat{J}_{k+1}(f_k(x_k, y_k, u_k, w_k) \mid x_k, y_k, u_k \right) \right\} \mid x_h \right]
\end{aligned}$$

1.22.3 Inventory Control (Section 3.2 of Bertsekas [2012a])

See Example 1.22.1 for setup.

Theorem 1.22.3.1 (**Order-up-to Polices Are Optimal**)**.** For each period k, there exists a threshold $S_k \geq 0$ such that

$$\mu_k^*(x_k) = [S_k - x_k]^+ = \begin{cases} S_k - x_k, & x_k \leq S_k, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. The DP algorithm is as follows:

$$\begin{aligned} J_N(x_N) &= 0, \\ J_k(x_k) &= \min_{u_k \geq 0} \left\{ c \cdot u_k + \mathop{\mathbb{E}}_{w_k} \left[\mathcal{G}(x_k + u_k - w_k) + J_{k+1}(x_k + u_k - w_k) \right] \right\} \end{aligned}$$

where again

$$\mathcal{G}(x_k + u_k - w_k) := \underbrace{p}_{\text{backorder cost/period}} \cdot \max\{0, w_k - x_k - u_k\} + \underbrace{i}_{\text{inventory cost/period}} \cdot \max\{x_k + u_k - w_k, 0\},$$

Note that $\mathcal{G}$ depends on k whenever the probability distribution of w_k depends on k, but we will assume all demands are identically distributed for simplicity (even though the analysis carries through easily if the distribution of demand is time-varying).

Now we will prove a lemma.

Lemma 1.22.3.2. $J_k(\cdot)$ is convex.

Proof. We will prove the result by induction. The base case is trivially true because $J_N(\cdot) = 0$ (note that we could replace $J_N(\cdot)$ with any convex function and the rest of the proof would follow). Assume $J_{k+1}(\cdot)$ is convex. Recall from Example 1.22.1

$$\begin{aligned} J_k(x_k) &= \min_{u_k \geq 0} \left\{ c \cdot u_k + \mathop{\mathbb{E}}_{w_k} \left[\mathcal{G}(\underbrace{x_k + u_k}_{y_k} - w_k) + J_{k+1}(x_k + u_k - w_k) \right] \right\} \\ &= \min_{y_k \geq x_k} \left\{ c \cdot (y_k - x_k) + \mathop{\mathbb{E}}_{w_k} \left[\mathcal{G}(y_k - w_k) + J_{k+1}(y_k - w_k) \right] \right\} \\ &= -cx_k + \min_{y \geq x_k} \left\{ cy + \mathop{\mathbb{E}}_{w_k} \left[\mathcal{G}(y - w_k) + J_{k+1}(y - w_k) \right] \right\} \\ &= -cx_k + \min_{y \geq x_k} \{ G_k(y) \} \end{aligned} \tag{1.37}$$

where we let $G_k(y) := cy + \mathop{\mathbb{E}}_{w_k}[\mathcal{G}(y - w_k) + J_{k+1}(y - w_k)]$. We claim that $G_k(y)$ is convex in y by the following argument. First, note that $\mathcal{G}(x) := p \cdot \max\{0, -x\} + i \max\{x, 0\}$ is convex in x. Therefore $\mathop{\mathbb{E}}_{w_k}[\mathcal{G}(y - w_k)]$ is convex in y since it is convex for each w_k and taking expectations preserves convexity.

Next, for each realized value of $w_k = \overline{w}_k$, $y \mapsto \mathcal{G}(y - \overline{w}_k) + J_{k+1}(y - \overline{w}_k)$ is convex because J_{k+1} is a convex function of an affine function and likewise with $\mathcal{G}(y - \overline{w}_k)$ (see Figure 1.5). Therefore the result follows when you take expectations.

We will briefly take a detour to consider the end behavior of $G_k(\cdot)$. As $y \to \infty$, we know that $cy \to \infty$. We know that $\mathcal{G}(\cdot)$ is always nonnegative and likewise with $J_{k+1}(y - w_k)$, so we must have that as $y \to \infty$, $G_k(\cdot) \to \infty$. As $y \to -\infty$, because of the assumption that $p > c$ (and that J_{k+1} is convex), we have that $G_k(y) \to \infty$. Therefore since $G_k(\cdot)$ is continuous, a minimizer of $G_k(\cdot)$ exists.

Let S_k be a minimizer of $G_k(\cdot)$. Note that if $x_k \geq S_k$, then $\min_{y \geq x_k} G_k(y) = G_k(x_k)$ due to the convexity of $G_k(\cdot)$. Clearly if $x_k \leq S_k$, then $\min_{y \geq x_k} G_k(y) = G_k(S_k)$. (Note that if S_k is not the only minimizer, this still holds.) In summary, we have

$$J_k(x_k) = \begin{cases} -cx_k + G_k(S_k) & x_k \leq S_k, \\ -cx_k + G_k(x_k) & x_k > S_k. \end{cases} \tag{1.38}$$

Each of these functions is continuous and convex in y. Therefore $J_k(x_k) = -cx_k + \min_{y \geq x_k} G_k(y)$ is convex in x_k. (See Figure 1.6.)

□

We have shown that a minimizer for G_k is given by S_k if $x_k < S_k$ and x_k otherwise. Then since $u_k = y_k - x_k$ in (1.37), a minimizer for (1.37) is attained at $u_k = S_k - x_k$ if $x_k < S_k$, and at $u_k = 0$ otherwise.

□

Next we will consider fixed orders cost. Before we find the optimal policy in this case, we will define some terms and derive some results we will need.

Definition 1.22.4 (**K-convexity; Definition 3.2.1 in Section 3.2 of Bertsekas [2012a], p. 130**). A real-valued function g is **K-convex** for $K \geq 0$ if

$$K + g(z + y) \geq g(y) + z\left(\frac{g(y) - g(y - b)}{b}\right) \tag{1.39}$$

for all $z \geq 0$, $b > 0$, and all $y \in \mathbb{R}$. Equivalent: for all $x < y < z'$,

$$K + g(z') \geq g(y) + (z' - y)\left(\frac{g(y) - g(x)}{y - x}\right). \tag{1.40}$$

(This construction follows from (1.39) by setting $x := y - b < y$ and $z' := z + y \geq y$.)

See Figure 1.7 for an example of a function that is K-convex but not convex.

Intuitively, from (1.40) we can think of this as meaning that the linear approximation of g at $z' = z + y$ by a secant line between y and $x = y - b$ is no more than K greater than $g(z + y)$. To make further sense of this definition, observe from (1.39) that if $K = 0$

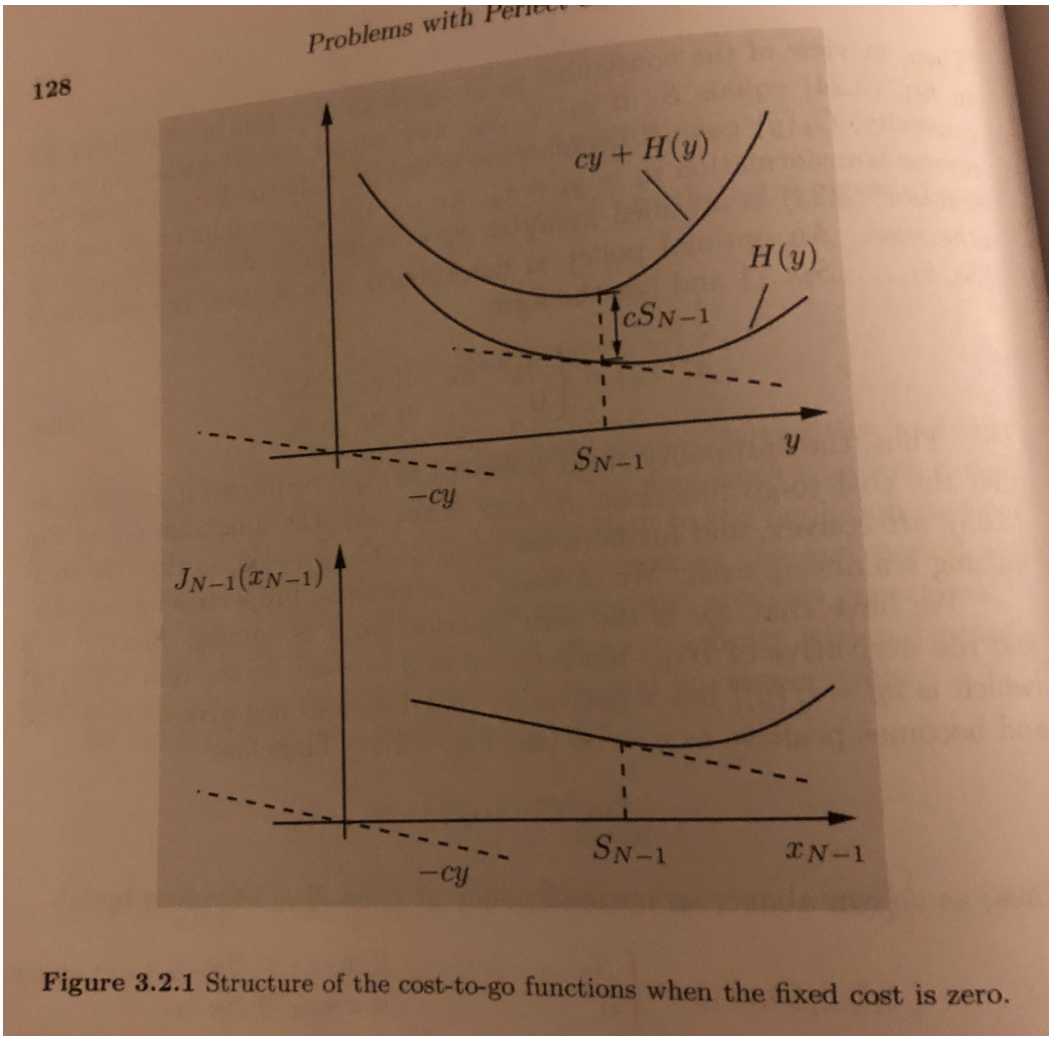

Figure 1.6: Cost-to-go function for inventory control problem. (The notation from Bertsekas [2012a] differs slightly from ours, but the lower figure shows that the cost-to-go is linear to the left of the minimizer and convex and nondecreasing to the right of the minimizer, as shown in Equation (1.38)).

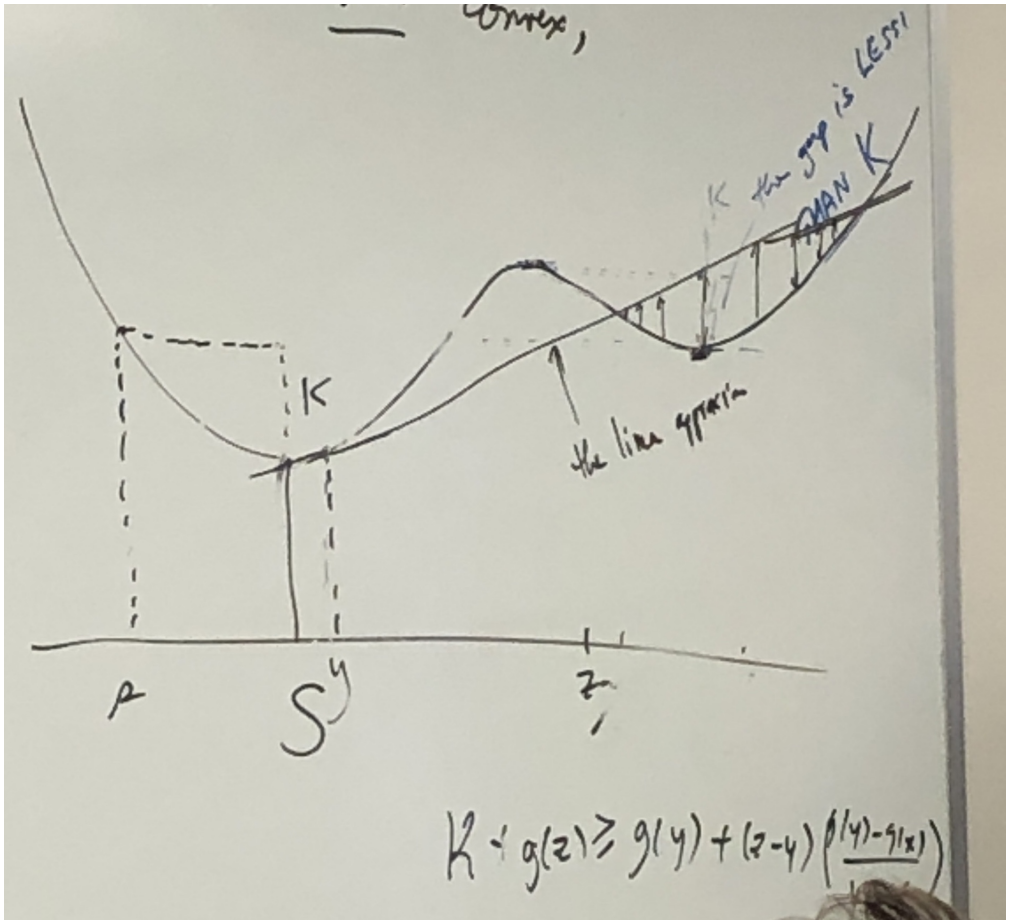

Figure 1.7: Function is K-convex (see Definition 1.22.4), but not convex.

$$\begin{aligned} & g(z+y) \geq g(y) + z\left(\frac{g(y)-g(y-b)}{b}\right) \\ \iff \quad & \frac{g(z+y)-g(y)}{z} \geq \frac{g(y)-g(y-b)}{b}; \end{aligned}$$

that is, the slopes of the secant lines are nondecreasing, which we know from Lemma **??** is true of convex functions. For the general case we have

$$\frac{K+g(z+y)-g(y)}{z} \geq \frac{g(y)-g(y-b)}{b}.$$

Next we will show some results about K-convex functions that will be useful.

Proposition 1.22.3.3 (**Lemma 3.2.1 in Bertsekas [2012a]**)**.** Properties of K-convex functions:

(a) A convex function g is 0-convex and K-convex for all $K \geq 0$.

(b) If g_1 is K-convex and g_2 is L-convex, then $\alpha g_1 + \beta g_2$ is $(\alpha K + \beta L)$-convex for all $\alpha > 0$, $\beta > 0$.

(c) If g is K-convex and w is a random variable, then $y \mapsto \mathbb{E}_w[g(y-w)]$ is also K-convex if $\mathbb{E}_w|g(y-w)| < \infty$ for all y. (Proof: exercise)

(d) If g is a continuous K-convex function with $g(y) \to \infty$ as $|y| \to \infty$, then there exist scalars (s, S) with $s \leq S$ such that

  (1) $g(S) \leq g(y)$ for all y,

  (2) $g(S) + K = g(s) \leq g(y)$ for all $y < s$. (See Figure 1.8.)

  (3) $g(y)$ is decreasing on $(-\infty, s)$.

  (4) $g(y) \leq g(z) + K$ for all y, z with $s \leq y \leq z$.

Proof of part (d). (1) Since g is continuous and $\lim_{|y|\to\infty} g(y) = \infty$, there exists a minimizing point S of g with minimum $g(S)$.

(2) Define $s := \min\{z : g(z) = g(S) + K\}$. Note that $s \leq S$. For all $y < s \leq S$, by K-convexity of g and the definition of s, from (1.40) we have

$$\begin{aligned} & K + g(S) \geq g(s) + (S-s)\left(\frac{g(s)-g(y)}{s-y}\right) \\ \iff & 0 \geq \underbrace{\frac{S-s}{s-y}}_{\geq 0} \cdot [g(s) - g(y)] \\ \iff & g(y) \geq g(s). \end{aligned}$$

(3) For any y_1, y_2 satisfying $y_1 < y_2 < s \leq S$, by K-convexity from (1.40) we have

$$K + g(S) \geq g(y_2) + \frac{S-y_2}{y_2-y_1}(g(y_2) - g(y_1)).$$

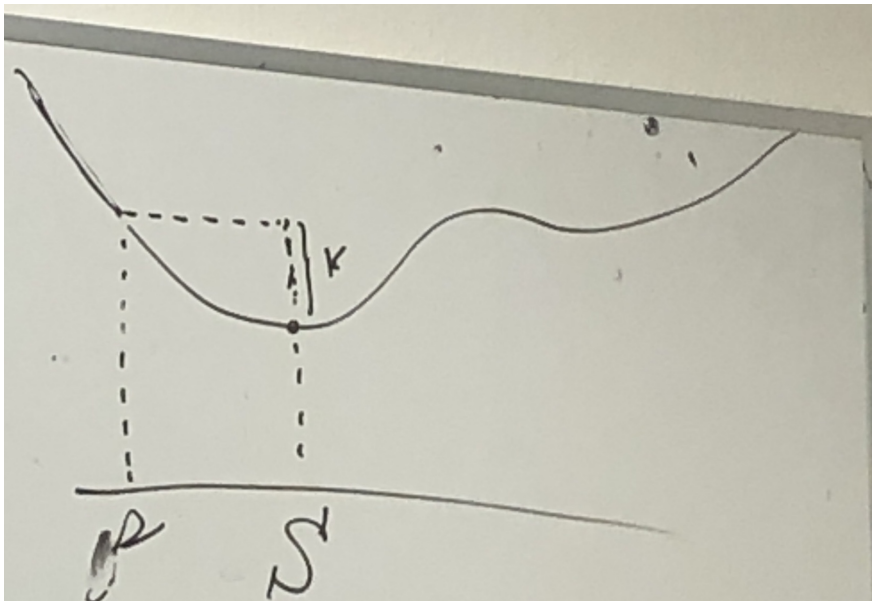

Figure 1.8: default

By (2) above, $g(y_2) > g(s) = g(S) + K$. Therefore

$$0 > g(S) + K - g(y_2) \geq \underbrace{\frac{S - y_2}{y_2 - y_1}}_{>0}(g(y_2) - g(y_1)), \qquad \implies g(y_1) > g(y_2).$$

(4) Note that $g(y) \leq g(z) + K$ holds when $y = z$ (trivially), $y = S$ (because S is a global minimizer of g), or $y = s$ (because $g(S) + K = g(s) \leq g(z) + K$ for all z). Since $s \leq y \leq z$, we only need to consider two cases.

- **Case 1:** Assume $S < y < z$. By K-convexity from (1.40),

$$\begin{aligned} K + g(z) &\geq g(y) + \underbrace{(z - y)}_{>0 \text{ (assumption)}} \cdot \left(\frac{\overbrace{g(y) - g(S)}^{>0 \ (S \text{ global min})}}{\underbrace{y - S}_{>0 \text{ (assumption)}}} \right) \\ &\geq g(y). \end{aligned}$$

- **Case 2:** Assume $s < y < S$. By K-convexity from (1.40),

$$\begin{aligned} & g(s) = K + g(S) \geq g(y) + (S - y)\left(\frac{g(y) - g(s)}{y - s}\right) \\ \iff \quad & \left(1 - \frac{\overbrace{y - S}^{>0}}{\underbrace{y - s}_{>y-S}}\right) g(s) \geq \left(1 - \frac{y - S}{y - s}\right) g(y) \\ \iff \quad & g(s) \geq g(y). \end{aligned}$$

Then $g(y) \leq g(s) = g(S) + K \leq g(z) + K$.

□

We are now ready to analyze inventory control with fixed costs. Suppose that

$$C(u) = \begin{cases} K + cu, & u > 0, \\ 0, & \text{otherwise} \end{cases}$$

so k is the fixed cost.

Theorem 1.22.3.4. When there is a fixed cost of order, an (s, S) policy is optimal; that is, for any period k there exists a pair of thresholds (s_k, S_k) with $s_k \leq S_k$ such that

$$\mu_k^*(x_k) = \begin{cases} S_k - x_k, & x_k \leq s_k, \\ 0, & \text{otherwise.} \end{cases} \tag{1.41}$$

Proof. Again let $y := x_k + u_k$. Define $G_k : \mathbb{R} \to \mathbb{R}$ as follows;

$$G_k(y) := cy + \mathop{\mathbb{E}}_{w_k}[\mathcal{G}(y - w_k) + J_{k+1}(y - w_k)]. \tag{1.42}$$

Then

$$\begin{aligned} J_N(x_N) &= 0, \\ J_k(x_k) &= \min_{u_k > 0} \{C(u_k) + \mathbb{E}\left[\mathcal{G}(x_k + u_k)\right] + \mathbb{E}\left[J_{k+1}(x_k + u_k - w_k)\right]\} \\ &= \min\left\{G_k(x_k) - cx_k, \min_{u_k > 0}\{K + G_k(x + u_k) - cx_k\}\right\} \\ &= -cx_k + \min\left\{G_k(x_k), \min_{y > x_k}[k + G_k(y)]\right\}, \qquad k \in \{0, \ldots, N = 1\}. \end{aligned}$$

where the third line follows because the cost is the minimum of the cost if no order is made (the cost on the left) or the cost if an order is made (the cost on the right), and the fourth line follows by the change of variable $y := x_k + u_k$.

If $G_k(y)$ is convex in y, the result follows quickly, as in the no fixed costs case covered in Theorem 1.22.3.1 (see Figure 1.9). In particular, the policy (1.41) is optimal, where S_k is a minimizer of $G_k(\cdot)$ and s_k is the smallest value of y for which $G_k(y) = K + G_k(S_k)$.

However, when $K > 0$ it is not necessarily true that G_k is convex. This means it could have multiple minima. Consider Figure 1.10. If it were true that G_k had a form like this, the optimal policy would be to order $(S - x)$ in interval I, zero in intervals II and IV, and $(\tilde{S} - x)$ in interval III.

But it turns out $G_k(\cdot)$ is K-convex, so the form of G_k in Figure 1.10 is impossible. If y_0 is the local maximum in in the interval III, we must have for sufficiently small $b > 0$

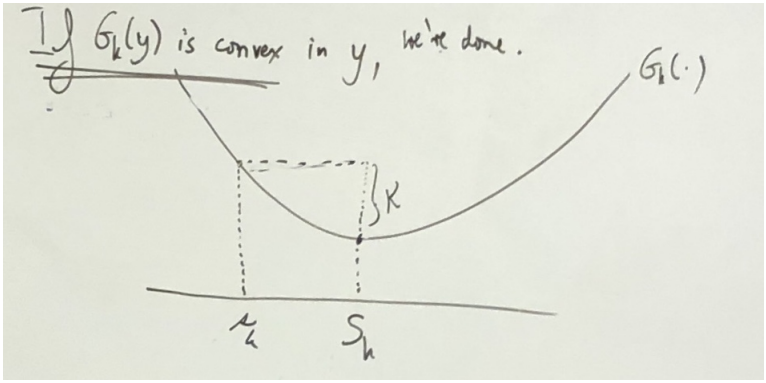

Figure 1.9: If $G_k(\cdot)$ is convex, the policy (1.41) is optimal, where S_k is a minimizer of $G_k(\cdot)$ and s_k is the smallest value of y for which $G_k(y) = K + G_k(S_k)$. However, G_k is not necessarily convex.

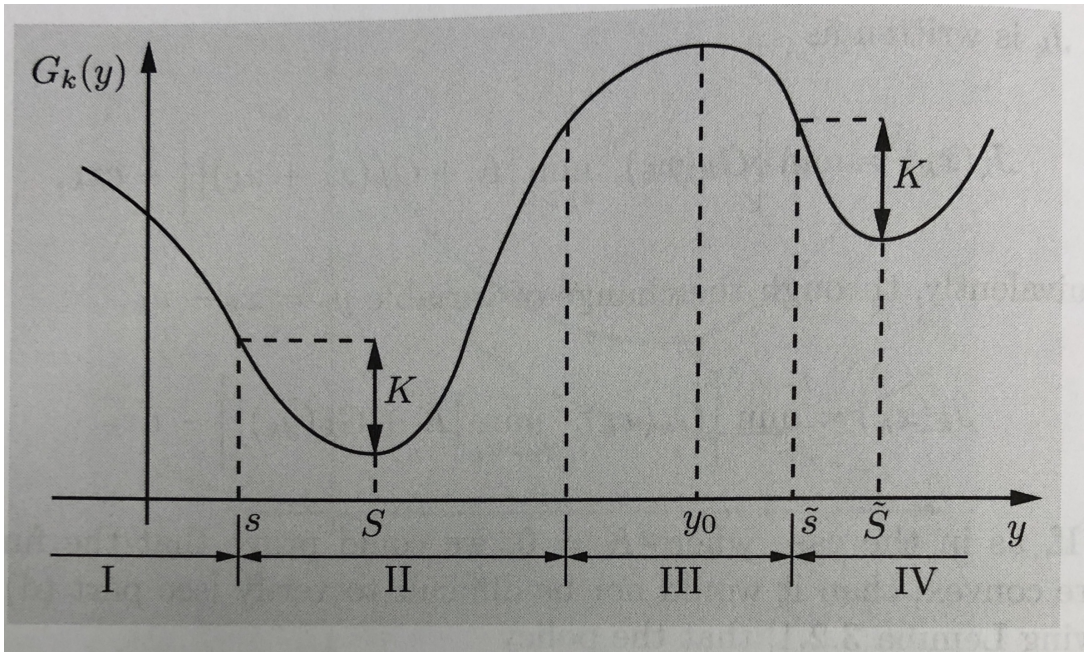

Figure 1.10: Figure 3.2.2 in Bertsekas [2012a].

$$\frac{G_k(y_0) - G_k(y_0 - b)}{b} \geq 0,$$

and from the definition of K-convexity (1.39) it follows that

$$K + G_k(\tilde{S}) \geq G_k(y_0),$$

contradicting the construction shown in Figure 1.10. Then the result will follow from (d)(4) of Proposition 1.22.3.3, because it is optimal to order up to the minimum if you have less inventory than s_k, but just to the right of s_k the cost of ordering exceeds the savings from ordering, and part (d)(4) of Proposition 1.22.3.3 guarantees that the cost of ordering will never again go below the savings from ordering. So the result is immediate from Lemma 1.22.3.5.

□

Lemma 1.22.3.5. G_k are continuous and K-convex for all k, and $G_k(y) \to \infty$ as $|y| \to \infty$. Further, $J_k(\cdot)$ are continuous and K-convex for all k.

Proof of Lemma 1.22.3.5. We will prove the result by induction. The base case is trivial since $J_N(\cdot) = 0$. Suppose $J_{\ell+1}(\cdot)$ is K-convex. Then

$$G_\ell(y) = \underbrace{cy}_{0-\text{convex}} + \underbrace{\mathop{\mathbb{E}}_{w_\ell} \mathcal{G}(y - w_l)}_{0-\text{convex}} + \underbrace{\mathop{\mathbb{E}}_{w_\ell} \left[J_{\ell+1}(y - w_\ell)\right]}_{K\text{-convex by (c) of Proposition 1.22.3.3}}$$

Therefore $G_\ell(y)$ is K-convex. Therefore by (d) of Proposition 1.22.3.3, there exist (s_ℓ, S_ℓ) such that

$$G_\ell(S_\ell) = \min_y G_\ell(y), \qquad G_\ell(s_\ell) = G_\ell(S_\ell) + K.$$

Then

$$\begin{aligned} J_\ell(x_\ell) &= -c \cdot x_\ell + \min\{G_\ell(x_\ell), \min_{y > x_\ell} [K + G_\ell(y)]\} \\ &= -cx_\ell + \begin{cases} \underbrace{K + G_\ell(S_\ell)}_{= G_\ell(s_\ell)}, & x_\ell \leq s_\ell, \\ G_\ell(x_\ell), & x_\ell > s_\ell \end{cases} \\ &= -cx_\ell + \begin{cases} G_\ell(s_\ell), & x_\ell \leq s_\ell, \\ G_\ell(x_\ell), & x_\ell > s_\ell. \end{cases} \end{aligned} \tag{1.43}$$

(See Figure 1.11 for an illustration of J_ℓ.) J_ℓ is continuous. Is $J_\ell(\cdot)$ convex? No. Consider the point s_ℓ. Left derivative of J_ℓ at s_ℓ is $-c + 0$, right derivative is $-c + \lim_{x \to s_\ell^+} G_\ell'(x) < -c + 0$ (because G_ℓ is still decreasing at s_ℓ—again see Figure 1.10). If the function were convex, the derivative would be nondecreasing, so this violates convexity.

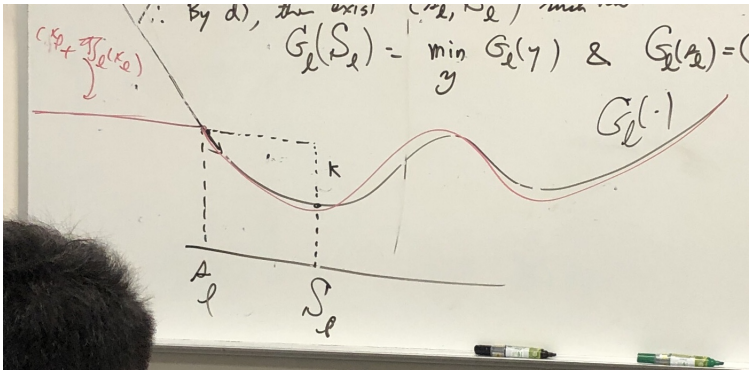

Figure 1.11: The red line is the right part of (1.43) ($J_\ell(x_\ell) + cx_\ell$).

However, we can show that J_ℓ is K-convex. We know from Equation (1.39) in Definition 1.22.4 that we must verify

$$K + J_\ell(y+z) \geq J_\ell(y) + z\left(\frac{J_\ell(y) - J_\ell(y-b)}{b}\right) \tag{1.44}$$

for all $y \in \mathbb{R}$, for all $z \geq 0, b \geq 0$. There are three cases to consider.

1. $y \geq s_\ell$: Then $y + z \geq s_\ell$. If $y - b \geq s_\ell$, then examining (1.43), the function $J_\ell(\cdot)$ at these points is $-cx_\ell + G_\ell(x_\ell)$ (the sum of a linear function and a K-convex function), so (1.44) holds and K-convexity follows. Suppose $y - b < s_\ell$. We need to verify (1.44); that is,

$$\begin{aligned}
& K + J_\ell(y+z) \geq J_\ell(y) + z\left(\frac{J_\ell(y) - J_\ell(y-b)}{b}\right) \\
\iff \text{(by (1.43))} \quad & K + \underbrace{G_\ell(y+z) - c(y+z)}_{=J_\ell(y+z),\ \text{since } y+z \geq s_\ell} \geq \underbrace{G_\ell(y) - cy}_{=J_\ell(y),\ \text{since } y \geq s_\ell} + z\left(\frac{\overbrace{G_\ell(y) - cy}^{=J_\ell(y),\ \text{since } y \geq s_\ell} - \overbrace{G_\ell(s_\ell) + c(y-b)}^{=J_\ell(y-b),\ \text{since } y-b < s_\ell}}{b}\right) \\
\iff \quad & K + G_\ell(y+z) - cz \geq G_\ell(y) + z\left(\frac{G_\ell(y) - G_\ell(s_\ell)}{b} - c\right) \\
\iff \quad & K + G_\ell(y+z) \geq G_\ell(y) + z\left(\frac{G_\ell(y) - G_\ell(s_\ell)}{b}\right), \qquad (1.45)
\end{aligned}$$

   so proving (1.45) is equivalent to proving K-convexity. We will look at two sub-cases:

   - **Sub-case 1:** $G_\ell(y) \geq G_\ell(s_\ell)$. By K-convexity of G_ℓ, using (1.40),

$$K + G_\ell(y+z) \geq G_\ell(y) + z\left(\frac{G_\ell(y) - G_\ell(s_\ell)}{y - s_\ell}\right) \geq G_\ell(y) + z\left(\frac{G_\ell(y) - G_\ell(s_\ell)}{b}\right)$$

     where the last step follows because $y - b < s_\ell \iff 1/(y - s_\ell) > 1/b$.

   - **Sub-case 2:** $G_\ell(y) < G_\ell(s_\ell)$. Then

$$\begin{aligned} K + G_\ell(y+z) &\geq K + G_\ell(S_\ell) \\ &= G_\ell(s_\ell) \\ &> G_\ell(y) && \text{(by assumption of sub-case)} \\ &\geq G_\ell(y) + z\left(\frac{\overbrace{G_\ell(y) - G_\ell(s_\ell)}^{<0}}{b}\right), \end{aligned}$$

verifying (1.45) and proving K-convexity in this case.

2. $y \leq y+z \leq s_\ell$: In this region, from (1.43) we see that J_ℓ is linear because $y - b \leq y \leq y+z \leq s_\ell$, so all of the J_ℓ in (1.44) are in the region where J_ℓ is constant, so K-convexity should hold trivially. Let's check: (1.44) becomes

$$K - c(y+z) + G_\ell(s_\ell) \geq -cy + G_\ell(s_\ell) + z\left(\frac{-cy + G_\ell(s_\ell) + c(y-b) - G_\ell(s_\ell)}{b}\right)$$
$$\iff \qquad K - cz \geq z\left(\frac{-cb}{b}\right),$$

so the inequality holds as desired, proving K-convexity in this case.

3. $y \leq s_\ell \leq y+z$: To establish K-convexity, from (1.44) we need to have

$$\begin{aligned} & K + J_\ell(y+z) \geq J_\ell(y) + \frac{z}{b}\left(J_\ell(y) - J_\ell(y-b)\right) \\ \iff\ & K + \underbrace{G_\ell(y+z) - c(y+z)}_{=J_\ell(y+z) \text{ since } y+z \geq s_\ell} \geq \underbrace{G_\ell(s_\ell) - cy}_{=J_\ell(y) \text{ since } y \leq s_\ell} + \frac{z}{b}\left(\underbrace{G_\ell(s_\ell) - cy}_{=J_\ell(y) \text{ since } y \leq s_\ell} - \underbrace{G_\ell(s_\ell) + c(y-b)}_{=J_\ell(y-b) \text{ since } y-b \leq s_\ell}\right) \\ \iff\ & K + G_\ell(y+z) - c(y+z) \geq G_\ell(s_\ell) - cy - cz \\ \iff\ & K + G_\ell(y+z) \geq G_\ell(s_\ell) \end{aligned} \tag{1.46}$$

But (1.46) is always true because $G_\ell(s_\ell) = G_\ell(S_\ell) + K \leq G_\ell(y+z) + K$ (since $G_\ell(S_\ell)$ is a global minimum), proving K-convexity in this case.

Therefore we have established K-convexity (and continuity) of both $G_\ell(\cdot)$ and $J_\ell(\cdot)$. It also holds that $G_\ell(y) \to \infty$ as $|y| \to \infty$. Now recall (1.42): this shows that $G_{\ell-1}(\cdot)$ is K-convex, since it is the sum of something linear, something K-convex (by Proposition 1.22.3.3 (c)), and something K-convex (as we just showed). But since $w_{\ell-1}$ is bounded, $G_{\ell-1}$ is continuous and $G_{\ell-1}(y) \to \infty$ as $|y| \to \infty$, so by the previous argument $J_{\ell-1}$ is K-convex. We can repeat this continually to show the result.

□

Remark 9. Also proved this in a 2000 paper "A single-unit approach to multi..." but the proof technique does not generalize to other problems as well (the way this proof does).

1.22.4 Capacity Allocation and Revenue Management

3 levels of revenue management:

1. Strategic: how to segment different markets and differentiate prices; which markets you go after, which subsets of customers, etc. (often done quarterly/annually.)
2. Tactical: (e.g. plane tickets: what's a good pricing strategy to maximize revenue? maybe save tickets for last minute when you can charge a lot?) calculate and update booking limits (done daily/weekly). Use forecasting, optimization, etc.
3. Booking Control/Execution: determine which booking to accept (done in real time)

Today: focus on tactical. 3 components:

1. **Resources:** units of capacities managed by suppliers. Example: seats on a flight leg, hotel room nights, rental car days. (We're interested in setting when resources are constrained.)
2. **Products:** what customers purchase. Each product corresponds to a combination of resources, for example, a ticket from LAX to JFK, may involve a layover (so 2 legs/resources), etc.
3. **Fare classes:** there are fares associated with end product (not just price, also things like layover time, bringing a carry-on, etc.). Each fare class is a combination of a price and restrictions on what/who can purchase a product.

A supplier controls

1. Sets of resources with a fixed and perishable capacity (e.g. plane seats perishable since can't sell tickets for seats on flights in the past).
2. A portfolio of products to offer to customers.
3. A set of fares associated with the end product. (has to do with setting prices, but main goal: the decision to open/close certain fare classes at each moment to maximize revenue.) Revenue: fare that you pay minus commission, etc.

Three problems in tactical revenue management:

1. **Single resource capacity allocation:** how many customers from different fare classes should we allow to book? ("single resource" e.g. direct flight or single night stay in hotel.)
2. **Network revenue management:** how should bookings be managed across a network of resources? (no known exact solution yet. we'll discuss after midterms)
3. **Overbooking:** how to managed bookings when faced with uncertain future no-shows? (standard ad-hoc methods.)

Today we'll focus on single resource capacity allocation.

Example 1.22.3 (**Two-class capacity allocation (Littlewood's formula)**)**.** Suppose we have C seats on a plane flight leg on a given date. We have two fare classes: a discounted fair P_d and full fare P_f, with $P_f > P_d$. Suppose the discount fare demand is D_d (number of customers that will arrive with a discount fare) and the full fare demand is D_f. Suppose D_d and D_f are independent, and suppose D_d customers arrive before D_f. Observation: we want to reserve seats for full-fare passengers.

To start, look at marginal analysis. Suppose we have x seats remaining ($C - x$ sold). Focus on the xth seat. Should we open it to a discount passenger or should we sell it? Breakeven point is when $P_d = \mathbb{P}(D_f \geq x)P_f$.

If y^* denotes an **optimal protection level** (how many seats we want to reserve for high-demand customers), then $P_d \leq P_f\mathbb{P}(D_f \geq y^*)$ and $P_d > P_f\mathbb{P}(D_f \geq y^* + 1)$. (Will show this formally). (y^* is the largest number such that $P_d \leq P_f\mathbb{P}(D_f \geq y^*)$. So

$$\begin{aligned}
& P_f\mathbb{P}(D_f \geq y^* + 1) < P_d \leq P_f\mathbb{P}(D_f \geq y^*) \\
\iff \quad & \mathbb{P}(D_f \geq y^* + 1) < \frac{P_d}{P_f} \leq \mathbb{P}(D_f \geq y^*) \\
\iff \quad & 1 - \mathbb{P}(D_f \geq y^*) \leq 1 - \frac{P_d}{P_f} < 1 - \mathbb{P}(D_f \geq y^* + 1) \\
\iff \quad & \mathbb{P}(D_f < y^*) \leq 1 - \frac{P_d}{P_f} < \mathbb{P}(D_f < y^* + 1) \\
\implies \quad & y^* = F_f^{-1}(1 - P_d/P_f),
\end{aligned}$$

where F_f is the cdf of demand for full-fare passengers. ($1 - P_d/P_f$ quantile). This is known as **Littlewood's formula**. (Recall: $F^{-1}(\alpha) := \inf\{x : F(x) \geq \alpha\}$.) More formally, the expected revenue given a booking limit b (maximum number of seats sold to discount passengers) is

$$\begin{aligned}
R(b) &= P_d \mathop{\mathbb{E}}_{D_d}[\min\{b, D_d\}] + P_f \mathop{\mathbb{E}}_{D_f}[\min\{D_f, C - \min\{b, D_d\}\}] \\
&= P_d \mathop{\mathbb{E}}_{D_d}[\min\{b, D_d\}] + P_f \mathop{\mathbb{E}}_{D_f}[\min\{D_f, \max\{C - b, C - D_d\}\}]
\end{aligned}$$

Assume D_d and D_f are continuous (so we can take derivatives) and assume we can interchange derivatives and expectations. Then (using almost-everywhere differentiability of the min function)

$$\begin{aligned}
R'(b) &= P_d\mathbb{E}[\mathbb{1}\{b < D_d\}] - P_f\mathbb{E}[\mathbb{1}\{b < D_d\} \cdot \mathbb{1}\{D_f > C - b\}] \\
&= P_d\mathbb{P}(b < D_d) - P_f\mathbb{P}(b < D_d) \cdot \mathbb{P}(D_f > C - b) \\
&= \mathbb{P}(b < D_d)\left[P_d - P_f\mathbb{P}(D_f > C - b)\right] \\
\implies \operatorname{sgn}(R'(b)) &= \operatorname{sgn}\left(\underbrace{P_d - P_f\mathbb{P}(D_f > C - b)}_{\text{decreasing in } b}\right).
\end{aligned}$$

Set equal to 0 to yield $P_d = P_f\mathbb{P}(D_f > C - b^*) \iff C - b^* = F_f^{-1}(1 - P_d/P_f)$. See Figure 1.12.

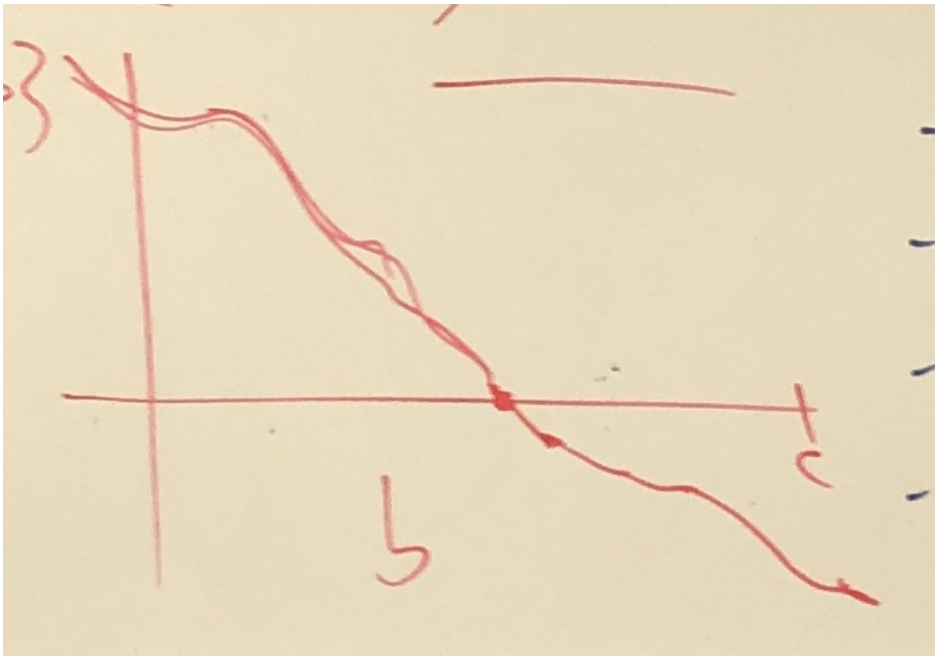

Figure 1.12: Figure for example 1.22.3.

Example 1.22.4 (**n-class capacity allocation**)**.** n fare classes: $P_1 > P_2 > \ldots > P_n$. Demand for class i: D_i, all independent. Sequential arrival of demand. Look at as DP. Let x be the number of seats remaining. Let $V_j(x)$ be the max expected seat that can be offered for class j, $j-1$, $j-2$, …, 2, 1, given that we have x seats remaining at the beginning of state j. Goal: compute $V_n(C)$. Initialization: $V_1(x) = P_1 \mathbb{E} \min\{x, D_1\}$,

$$V_j(x) = \max_{y_{j-1} \leq x} \left\{ \underbrace{P_j \mathbb{E}\left[\min\{D_j, x - y_{j-1}\}\right] + \mathbb{E}\left[V_{j+1}\left(\max\{y_{j-1}, x - D_j\}\right)\right]}_{W_j(x, y_{j-1})} \right\}$$

where y_{j-1} is the amount of seats reserved for class $j-1, j-2, \ldots, 2, 1$.

Theorem 1.22.4.1. For the capacity allocation problem with C seats (Example 1.22.4), the optimal booking control policy is given a nested protection level $y_1^* \leq \ldots \leq y_n^* = C$. (e.g., y_2^* is the number of seats reserved for class 1 and 2, y_3 is left for 3, 2, and 1, etc. Fixed number independent of remaining capacity.)

For ease of exposition, assume demand is continuous (so we can take derivatives. can just use finite differences if you want to be discrete, just a pain.). First we will need a lemma.

Lemma 1.22.4.2. Suppose demand and inventory are continuous. Then for each $j \geq 1$,

1. $V_j(\cdot)$ is an increasing and concave function.
2. There exist optimal protection levels y_j^* for $j \in [n]$ that maximize expected revenue, given by the solutions to $V'_{j-1}(y_{j-1}^*) = P_j$.

3. $V_j'(x) \geq V_{j-1}'(x)$ for all x.

Proof. We'll prove that $V_j(\cdot)$ is an increasing and concave function by induction on j. For $j = 1$, $V_1(x) = P_1\mathbb{E}[\min\{x, D_1\}]$, so it's trivially true. Next, assume that the result holds for $V_{j-1}(\cdot)$. We'll prove that it's true for $V_j(\cdot)$. For all $x \geq y$,

$$W_j(x,y) = P_j\mathbb{E}\left[\min\{D_j, x-y\}\right] + \mathbb{E}\left[V_{j+1}\left(\max\{y, x-D_j\}\right)\right], \qquad V_j(x) = \min_{y \leq x} W_j(x,y).$$

Then

$$\begin{aligned}\frac{\partial}{\partial y} W_j(x,y) &= -P_j\mathbb{E}\left[\mathbb{1}\{D_j > x-y\}\right] + \mathbb{E}\left[V_{j-1}'(y)\cdot\mathbb{1}\{x-D_j < y\}\right] \\ &= \mathbb{P}(D_j > x-y)(-P_j + V_{j-1}'(y)).\end{aligned}$$

By the inductive hypothesis, $V_{j-1}(\cdot)$ is concave, so $-P_j + V_{j-1}'(y)$ is decreasing in y. Notice that $V_{j-1}'(0) \geq P_{j-1} > P_j$ since $V_{j-1}'(0)$ is the marginal revenue. Therefore there exists y_{j-1}^* such that for $y < y_{j-1}^*$ $\frac{\partial}{\partial y} W_j(x,y)$ is positive and for $y > y_{j-1}^*$ $\frac{\partial}{\partial y} W_j(x,y)$ is negative. Then

$$V_j(x) = \max_{y \leq x} W_j(x,y) = W_j(x, \min\{x, y_{j-1}^*\})$$

Then the optimal protection level y_{j-1}^* for class $j-1, j-2, \ldots, 2, 1$ is the value such that

1. For continuous demand: $V_{j-1}'(y_{j-1}^*) = P_j$. (notice that this is the same as in the two-class case.)
2. For discrete demand **(need for Poisson case in homework 2 question 6)**: $y_{j-1}^* = \max\{y : P_j < \Delta V_{j-1}(y)\}$ where $\Delta V_{j-1}(y) = V_{j-1}(y) - V_{j-1}(y-1)$.

Consider the continuous case.

$$\begin{aligned}V_j(x) &= W_j(x, \min\{x, y_{j-1}^*\}) && (1.47)\\ &= P_j\mathbb{E}[\min\{D_j, (x-y_{j-1}^*)^+\}] + \mathbb{E}[V_{j-1}(\max\{x-D_j, \min\{x, y_{j-1}^*\}\})] && (1.48)\\ &= \begin{cases} V_{j-1}(x), & x \leq y_{j-1}^* \\ P_j\mathbb{E}[\min\{D_j, x-y_{j-1}^*\}] + \mathbb{E}[V_{j-1}(\max\{x-D_j, y_{j-1}^*\})], & x \geq y_{j-1}^* \end{cases} && (1.49)\\ \implies V_j'(x) &= \begin{cases} V_{j-1}'(x), & x \leq y_{j-1}^* \\ P_j\mathbb{E}[\mathbb{1}\{D_j > x-y_{j-1}^*\}] + \mathbb{E}[V_{j-1}'(x-D_j)\mathbb{1}\{x-D_j > y_{j-1}^*\}], & x \geq y_{j-1}^* \end{cases} && (1.50)\end{aligned}$$

Goal: show $V_j'(x)$ is decreasing in x. We already know $V_{j-1}'(x)$ is decreasing in x by inductive hypothesis. So we focus on the other part (when $x \geq y_{j-1}^*$).

$$
\begin{aligned}
&P_j\mathbb{E}[\mathbb{1}\{D_j > x - y^*_{j-1}\}] + \mathbb{E}[V'_{j-1}(x - D_j)\mathbb{1}\{x - D_j > y^*_{j-1}\}] \\
&= P_j\mathbb{E}[1 - \mathbb{1}\{D_j < x - y^*_{j-1}\}] + \mathbb{E}[V'_{j-1}(x - D_j)\mathbb{1}\{x - D_j > y^*_{j-1}\}] \\
&= P_j + \mathbb{E}\left[\mathbb{1}\{D_j < x - y^*_{j-1}\}(-P_j + V'_{j-1}(x - D_j))\right] \qquad (1.51)
\end{aligned}
$$

Note for each D_j, the expression inside the expectation operator is nonincreasing in x (see Figure 1.13). So it holds over expectations that $V'_j(x)$ is nonincreasing in x. Now we want to show that $V'_j(x) \geq V'_{j-1}(x)$ for all x. Again, it holds trivially that if $x \leq y^*_{j-1}$ then $V'_j(x) = V'_{j-1}(x) \geq V'_{j-1}(x)$, so we focus on the case $x \geq y^*_{j-1}$. Note that $\mathbb{1}\{x - D_j \geq y^*_{j-1}\} \geq \mathbb{1}\{x \geq y^*_{j-1}\}$ and that since $V'_{j-1}(\cdot)$ is nonincreasing $V'_{j-1}(x - D_j) \geq V'_{j-1}(x)$. Then from (1.51)

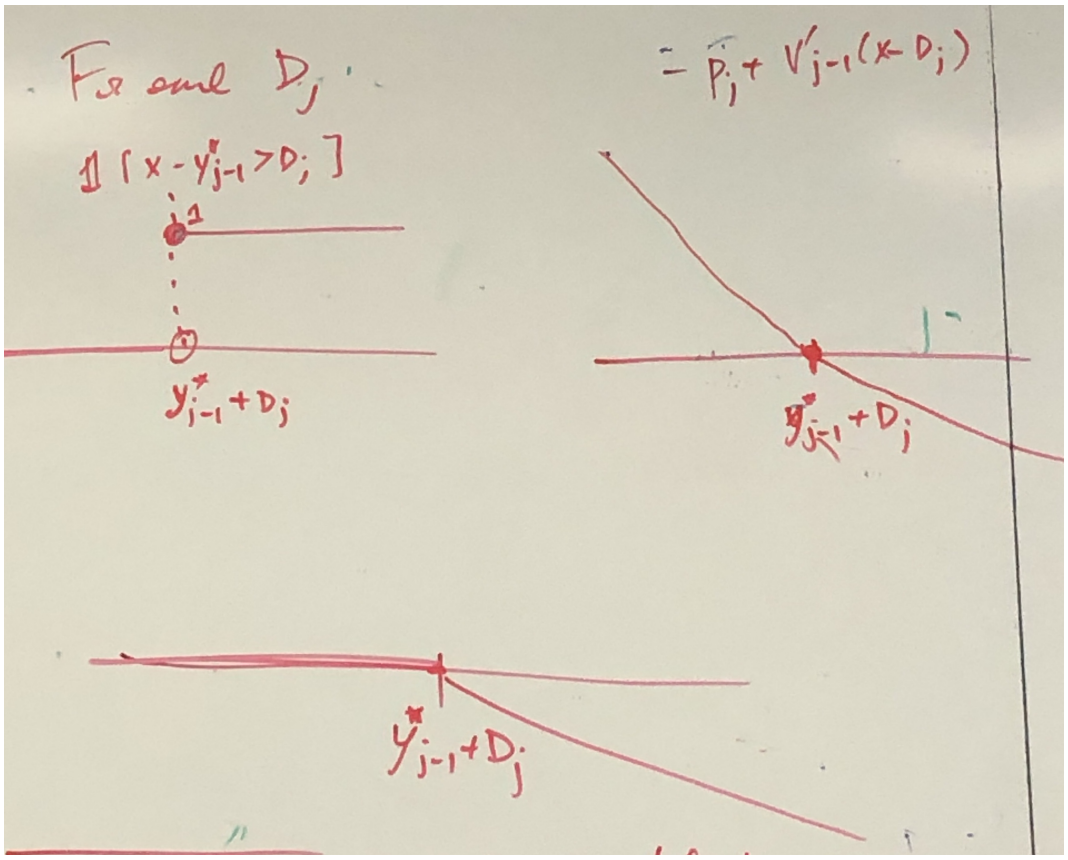

Figure 1.13: The argument of the expectation operator in (1.51) is nonincreasing in x regardless of the value of the random variable D_j, so the expectation itself is nonincreasing in x.

$$
\begin{aligned}
V'_j(x) &= P_j + \mathbb{E}\left[\mathbb{1}\{x - D_j \geq y^*_{j-1}\}(-P_j + V'_{j-1}(x - D_j))\right] \\
&\geq P_j + \mathbb{E}\left[\mathbb{1}\{x \geq y^*_{j-1}\}(-P_j + V'_{j-1}(x)\right] \\
&= V'_{j-1}(x).
\end{aligned}
$$

□

Now we can prove the main result.

Proof of Theorem 1.22.4.1. It only remains to show that the optimal protection levels are nested. We want to show that $y_j^* \geq y_{j-1}^*$. From Lemma 1.22.4.2 we know that the optimal protection levels are given by $V_j'(y_j^*) = P_{j+1}$. What does $V_j'(\cdot)$ look like? It's given in (1.50). Since $P_j > P_{j+1}$ and per Lemma 1.22.4.2 $V_j'(x) \geq V_{j-1}'(x)$ for all x, it holds that $y_{j-1}^* < y_j^*$. See a depiction in Figure 1.14.

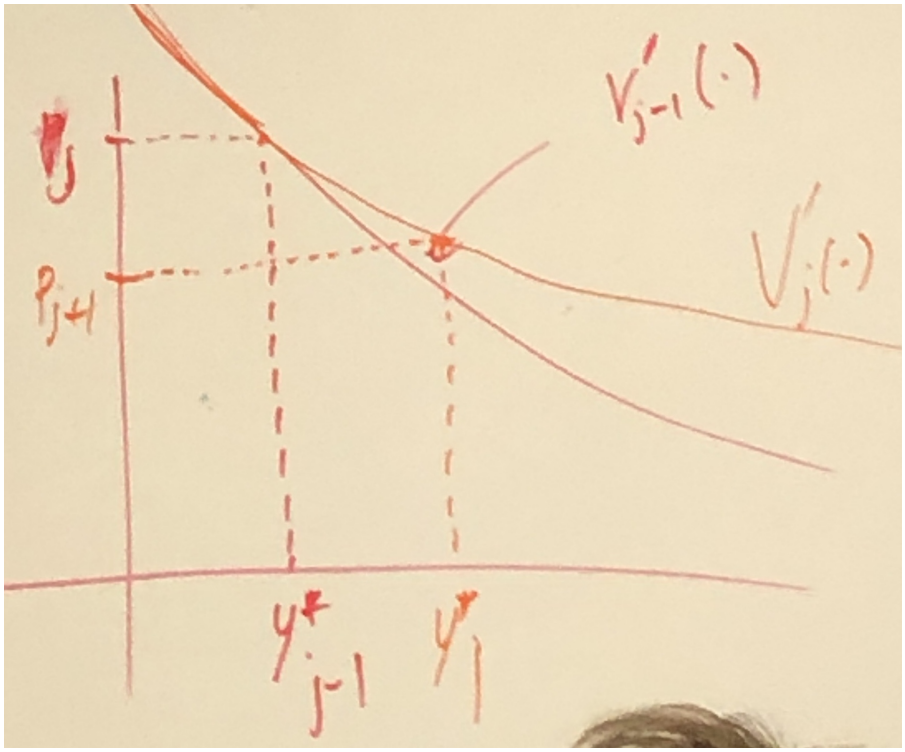

Figure 1.14: Illustration of nested protection levels from Theorem 1.22.4.1.

□

1.22.5 Optimal Stopping (Section 3.4 of Bertsekas [2012a])

Example 1.22.5 (**Optimal stopping**). You own an asset. You are given money $w_h \geq 0$ in period h. If you accept an offer, you can invest the money at a risk-free rate $r > 0$. If you reject the offer, you wait for the next one. Assume $w_0, w_1, \ldots, w_{N-1}$ are independent random variables. Goal: maximize. the total reward (money) at the end of period N. Now

$$J_k(x) = \max\left\{\underbrace{(1+r)^{N-k}x}_{\text{if you stop}}, \underbrace{\mathbb{E}\left[J_{k+1}(w_k)\right]}_{\text{if you don't stop}}\right\}, \qquad J_N(x) = x.$$

Optimal policy: stop (accept the current offer) if and only if

$$x \geq \frac{\mathbb{E}_{w_k}[J_{k+1}(w_k)]}{(1+r)^{N-k}} := \underbrace{\alpha_k}_{\text{threshold in period } k}.$$

Claim: If $w_0, w_1, \ldots, w_{N-1}$ are i.i.d., then $\alpha_0 \geq \alpha_1 \geq \ldots \geq \alpha_{N-1}$.

Let $V_k(x) := J_k(x)/(1+r)^{N-k}$. We have $V_N(x) = x$,

$$V_k(x) = \max\left\{x, \underbrace{\frac{\mathbb{E}[V_{k+1}(w)]}{1+r}}_{\text{state independent: no dependence on } x}\right\}.$$

We'll show that $V_0(\cdot) \geq V_1(\cdot) \geq \ldots \geq V_{N-1}(\cdot)$. Base case: $V_{N-1}(x) \geq V_N(x)$, obvious. Next:

Suppose $V_{k+1}(\cdot) \geq V_{k+2}(\cdot)$. Then

$$\begin{aligned} V_k(x) &= \max\left\{x, \frac{\mathbb{E}[V_{k+1}(w)]}{1+r}\right\} \\ &\geq \max\left\{x, \frac{\mathbb{E}[V_{k+2}(w)]}{1+r}\right\} \\ &= V_{k+1}(x). \end{aligned}$$

Example 1.22.6 (Exotic options, optimal stopping). Suppose that the price of your asset (option) depends on the prices of $n = 10$ other assets. So the state is $(x_1, \ldots, x_{10})$. Then

$$J_k(x_1, \ldots, x_{10}) = \max\left\{\underbrace{G(x_1, \ldots, x_{10})}_{\text{stop now}}, \underbrace{\mathbb{E}\left[J_{k+1}(\tilde{x}_1, \ldots, \tilde{x}_{10}\right]}_{\text{don't stop now}}\right\}$$

where $J_k(x_1, \ldots, x_{10})$ is the current prices of the 10 assets that influence/determine the value of your "option" or asset and $G(x_1, \ldots, x_{10})$ is the reward for selling, and

$$J_N(x_1, \ldots, x_N) = g(x_1, \ldots, x_{10}).$$

Suppose that x_i takes values $\{1, \ldots, 100\}$.

1.22.6 Infinite Horizon (Sections 1.2, 1.5 and 2.1 of Bertsekas [2012b]; starts on p. 210 of pdf for Volume 3)

Restatement of setup: infinite horizon problems with discounted cost:

$$x_{k+1} = f(x_k, u_k, w_k)$$

$$J^*(x_0) = \min_{\pi \in \Pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \mid x_0\right]$$

Stationary discrete-time systems

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k \in \{0, 1, \ldots\}$$

where for all k, $x_k \in \mathcal{X}$ is the state at time k, $u_k \in \mathcal{U}$ is the control, and we require $u_k \in \mathcal{U}(x_k)$ (that is, there is a set of allowable controls given state x_k, and $w_k \in \mathcal{W}$ are disturbances.

Important: the random disturbances $w_k : k \in \{0, 1, 2, \ldots\}$ are characterized by a transition probability $\mathbb{P}(\cdot \mid x_k, u_k)$ that is independent of k. (only dependent on state and control; not time. Nothing indexed by time; assume process is stationary.) Also, $w_k \perp\!\!\!\perp w_{k-1}, w_{k-2}, \ldots, w_0$.

Given an initial state x_0, we want to find a policy $\Pi = \{\mu_0, \mu_1, \mu_2, \ldots\}$ where $\mu_k : \mathcal{X} \to \mathcal{U}$ with $\mu_k(x_k) \in \mathcal{U}(x_k)$ that minimizes the cost for

$$J_\pi(x_0) := \limsup_{N \to \infty} \mathop{\mathbb{E}}_{w_0, w_1, \ldots}\left[\sum_{k=0}^{N} \alpha^k g(x_k, \mu_k(x_k), w_k)\right]$$

where $\alpha \in (0, 1)$ is the discount factor (typically close to 1).

Let Π be the space of admissible polices. Our goal is to determine an optimal cost function $J^* : \mathcal{X} \to \mathbb{R}$

$$J_k^* = \min_{\pi \in \Pi} J_\pi(x).$$

Key reason to study infinite horizon problems: in most cases, an optimal policy is stationary: $\Pi^* = \{\mu, \mu, \mu, \ldots\}$. If the policy is stationary, we write $J_\mu(\cdot)$ (a policy is a sequence of functions mapping states to actions, but often in this case the policy is stationary, so the sequence is just the same function repeated). A stationary policy is optimal if $J_\mu(x) = J^*(x)$ for all $x \in \mathcal{X}$.

Definition 1.22.5 (**MDP [Markov decision process] Assumption**)**.** We assume that the state space $\mathcal{X}$, the control set $\mathcal{U}$, and the disturbance space $\mathcal{W}$ are all finite.

The MDP assumption corresponds to the classical finite-state Markov decision process. Let $\mathcal{X} = \{1, \ldots, n\}$. The transition matrix is given by

$$P_{ij}(u) = \mathbb{P}(X_{k+1} = j \mid X_k = i, U_k = u).$$

We can write this using the system dynamics notation as follows:

$$x_{k+1} = f(x_k, u_k, w_k) = w_k$$

For each decision u,

$$W_{ij}(u) = \{w \in \mathcal{W} \mid f(i,u,w) = j\}, \qquad P_{ij}(u) = \mathbb{P}(W_{ij}(u) \mid i,u).$$

Example 1.22.7. $X = \{1,2\}$, $U = \{A,B\}$.

$$\mathbb{P}_{ij}(u=A) = \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{pmatrix}, \qquad \mathbb{P}_{ij}(u=B) = \begin{pmatrix} 1/6 & 5/6 \\ 5/6 & 1/6 \end{pmatrix}$$

Define a random disturbance W as follows (we need to specify $\mathbb{P}(W = j \mid i,u)$):

$$\mathbb{P}(W=j \mid i,u) = \begin{cases} 1/2 \text{ for } j=1, & 1/2 \text{ for } j=2 & \text{if } i=1, u=A \\ 3/4 \text{ for } j=1, & 1/4 \text{ for } j=2 & \text{if } i=2, u=A \\ 1/6 \text{ for } j=1, & 5/6 \text{ for } j=2 & \text{if } i=1, u=B \\ 5/6 \text{ for } j=1, & 1/6 \text{ for } j=2 & \text{if } i=2, u=B \end{cases}$$

Given the MDP assumption, 2 key questions.

1. Does the optimal cost J^* satisfy some kind of Bellman equation?
2. How do we compute the optimal policy?

Motivation for the Bellman Equations

$J^*(x)$ is the optimal cost function (stationary). Will satisfy (guess, will prove)

$$J^*(x) = \min_{u \in \mathcal{U}(x)} \{\mathbb{E}[g(x,u,w)] + \alpha \mathbb{E}\left[J^*(f(x,u,w))\right]\}$$

Consider an arbitrary policy $\pi = \{\mu_0, \mu_1, \mu_2, \ldots\}$. Suppose we consider the cumulative cost of the first n states and add some terminal cost $\alpha^n J(x_n)$. The expected total cost is

$$\mathop{\mathbb{E}}_{w_0,\ldots,w_{n-1}} \left[\alpha^n J(x_n) + \sum_{k=0}^{n-1} \alpha^k g(x_h, \mu_h(x_h), w_k) \right]$$

Suppose we want to find the minimum cost of N states $H_N(x), \alpha^N J(x)$. For $h \in \{1,2,\ldots,N\}$,

$$H_{n-h}(x) = \min_{u \in \mathcal{U}(x)} \mathop{\mathbb{E}}_{w} \left[\alpha^{N-h} g(x,u,w) + H_{n-h+1}(f(x,u,w)) \right].$$

$$\iff H_{n-h-1}(x) = \min_{u \in \mathcal{U}(x)} \mathop{\mathbb{E}}_{w} \left[\alpha^{N-h-1} g(x,u,w) + H_{n-h}(f(x,u,w)) \right].$$

let $V_k(x) := \frac{H_{n-k}(x)}{\alpha^{N-k}}$. Then we have the following DP algorithm:

$$V_0(x) = J(x); \qquad V_{k+1}(x) = \min_{u \in \mathcal{U}(x)} \mathop{\mathbb{E}}_{w} \left[g(x,u,w) + \alpha V_k(f(x,u,w0) \right]$$

Intuition: $V_k(x)$ is the minimum net present value of the cost of an h-stage problem given that we start in state X.

Definition 1.22.6. A dynamic programming operator $T : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$ is defined as follows: given $J : \mathcal{X} \to \mathbb{R}$, with $J \in \mathbb{R}^{|\mathcal{X}|}$, for all $x \in \mathcal{X}$, $TJ \in \mathbb{R}^{|\mathcal{X}|}$ where

$$(TJ)(x) := \min_{u \in \mathcal{U}(x)} \left\{ \mathop{\mathbb{E}}_{w} [g(x,u,w) + \alpha J(f(x,u,w))] \right\}$$

Important observation: TJ is the optimal cost for a one stage problem that has a state cost g and a terminal cost αJ.

Similarly, for any stationary policy $\mu : \mathcal{X} \to \mathcal{U}$ and $J : \mathcal{X} \to \mathbb{R}$ define $T_\mu : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$ such that for all $x \in \mathcal{X}$

$$(T_\mu J)(x) := \mathbb{E}\left[g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w)) \right]$$

Notation: $T^k = T \circ T \circ \ldots \circ T$. T_k is the optimal cost for the k-stage α-discounted problem with an initial state x, cost-per-state g, and terminal cost $\alpha^k J$.

Note; if $\Pi = \{\mu_0, \ldots, \mu_{k-1}\}$, then the cost of the policy Π for the k-stage problem with initial state x, cost-per-state g, and terminal cost $\alpha^k J$ is $(T_{\mu_0} T_{\mu_1} \cdots T_{u_{k-1}} J)(x)$.

One last remark: since typically $g(x,u,w) \geq 0$, we have the inequality

$$(TJ)(x) \leq \alpha J(f(x,u,w)). \tag{1.52}$$

Definition 1.22.7 (Contraction Mapping (Section 1.5 of Bertsekas [2012b])). Let $(Y, \|\cdot\|)$ be a real normed vector space. A function $T : Y \to Y$ is said to be a **contraction mapping** if. for some $\alpha \in (0,1)$, we have

$$\|Ty - Tz\| \leq \alpha \|y - z\|, \qquad \forall y \in Y.$$

The scalar α is said to be the **modulus of contraction** of T.

(See also Definition **??**. For more on contraction mappings, see Section **??**.)

If Y is a Banach space (that is, complete under the norm $\|\cdot\|$), a contraction mapping $F : Y \to Y$ has a unique fixed point; that is, the equation $y = Fy$ has a unique solution y^* called the **fixed point** of F. Further, the sequence $\{y_k\}$ generated by the iteration $y_{k+1} = Fy_k$ converges to y^* starting from an arbitrary initial point y_0.

We will want to show that the dynamic programming operator is a contraction mapping. It will be enough to show that the dynamic programming operator is **monotonic** and has **constant shift.**

Theorem 1.22.6.1 (Blackwell's Sufficient Conditions). Let $\mathbb{R}^{|X|}$ be a space of bounded functions $J : X \to \mathbb{R}$ with the sup metric. Let $T : \mathbb{R}^{|X|} \to \mathbb{R}^{|X|}$ be an operator satisfying two conditions:

1. **(Monotonicity)** If $J, J' \in \mathbb{R}^{|X|}$ and $J(x) \leq J'(x)$ for all $x \in X$, then $(TJ)(x) \leq (TJ')(x)$ for all $x \in X$.

2. **(Discounting)** There exists $\alpha \in (0,1)$ such that

$$(T(J+c))(x) \leq (TJ)(x) + \alpha c \qquad \forall J \in \mathbb{R}^{|X|}, c \geq 0, x \in X.$$

Then T is a contraction mapping with modulus α.

Proof. Let $J, J' \in \mathbb{R}^{|X|}$. Then

$$J(x) = J'(x) + J(x) - J'(x) \leq J'(x) + \sup_{y \in X} |J(y) - J'(y)| = J'(x) + \|J - J'\|_\infty, \qquad \forall x \in X,$$

which we can write in shorthand as

$$J \leq J' + \|J - J'\|_\infty.$$

Properties (1) and (2) imply that for some $\alpha \in (0,1)$

$$\begin{aligned} TJ \leq T(J' + \|J - J'\|_\infty) \leq TJ' + \alpha \|J - J'\|_\infty \qquad &\iff TJ - TJ' \leq \alpha \|J - J'\|_\infty, \\ TJ' \leq T(J + \|J - J'\|_\infty) \leq TJ + \alpha \|J - J'\|_\infty \qquad &\iff TJ' - TJ \leq \alpha \|J - J'\|_\infty. \end{aligned}$$

Combining these, we have

$$\begin{aligned} |(TJ)(x) - (TJ')(x)| \leq \alpha \|J - J'\|_\infty \qquad &\forall x \in X \\ \implies \qquad \sup_{x \in X} |(TJ)(x) - (TJ')(x)| \leq \alpha \|J - J'\|_\infty \\ \implies \|TJ - TJ'\|_\infty \leq \alpha \|J - J'\|_\infty. \end{aligned}$$

□

Now we will show that these properties hold.

Theorem 1.22.6.2 (Properties of T and T_μ). 1. **Monotonicity: (Lemma 1.1.1 in Bertsekas [2012b], p.9)** for any $J : \mathcal{X} \to \mathbb{R}$ and $J' : \mathcal{X} \to \mathbb{R}$ such that $J(x) \leq J'(x)$ for all $x \in \mathcal{X}$, then

$$(T^k J)(x) \leq (T^k J')(x), \qquad (T_\mu^k J)(x) \leq (T_\mu^k J')(x).$$

for any $k \geq 0$.

2. **Constant shift (Lemma 1.1.2 in Bertsekas [2012b], p. 9):** For each $k \geq 0$, $J : \mathcal{X} \to \mathbb{R}$ and scalar r,

$$[T^k(J + r\boldsymbol{e})](x) = (T^k J)(x) + \alpha^k r, \qquad [T_\mu^k(J + r\boldsymbol{e})](x) = (T_\mu^k J)(x) + \alpha^k r,$$

where $\boldsymbol{e}$ is a vector of all 1s.

Proof of monotonicity (Lemma 1.1.1 in Bertsekas [2012b]). □

Proof of constant shift (Lemma 1.1.2 in Bertsekas [2012b]). □

Now we can show that the dynamic programming operator is a contraction mapping.

Theorem 1.22.6.3 (**Proposition 1.2.6 in Bertsekas [2012b]**)**.** The dynamic programming operator T is a contraction mapping in the normed vector space $(\mathbb{R}^{|X|}, \|\cdot\|_\infty)$ (where $\|J\|_\infty = \sup_{x \in \mathcal{X}} |J(x)|$).

Proof. Theorem 1.22.6.2 shows that the sufficient conditions of Theorem 1.22.6.1 are satisfied for the dynamic programming operator T.

□

(For more on contraction mappings, see Section **??**.)

3 questions to prove (similar to Proposition 7.2.1, p. 214 of pdf of 3rd edition):

1. **Convergence of DP algorithm (second part of Proposition 1.5.1 in Bertsekas [2012b], p. 48; second part of Proposition 1.5.4, p. 53):** under the MDP assumption, for every $J : \mathcal{X} \to \mathbb{R}$ and for any policy $\mu : \mathcal{X} \to \mathcal{U}$, $J^* = \lim_{k \to \infty} T^k J$, $\quad J_\mu = \lim_{h \to \infty} T_\mu^h J$.

   note that

$$(TJ^*)(x) = \min_{u \in \mathcal{U}(x)} \{\mathbb{E}\left[g(x, u, w) + \alpha J^*(f(x, u, w))\right]\} = J^*(x).$$

   Statement in book: If $T : \mathbb{R}^{|X|} \to \mathbb{R}^{|X|}$ is a contraction mapping with modulus $\alpha \in (0, 1)$, then $\{T^k J\}$ converges to J^* for any $J \in \mathbb{R}^{|X|}$, and we have

$$\|T^k J - J^*\| \leq \alpha^k \|J - J^*\|, \qquad k \in \mathbb{N}.$$

   notes from review on 2/11/10: J^* is the unique fixed point of T and J_μ is the unique fixed point of T_μ and $J^* = \lim_{k \to \infty} T^k J$ (used contraction mapping to prove).

2. **Bellman's equations (first part of Proposition 1.5.1 in Bertsekas [2012b], p. 48; first part of Proposition 1.5.4, p. 53).** The optimal cost J^* is the unique solution to the equation

$$J^*(x) = \min_{u \in \mathcal{U}(x)} \{\mathbb{E}\left[g(x, u, w) + \alpha J^*(f(x, u, w))\right]\} \tag{1.53}$$

   That is, $J^* = TJ^*$. (Statement in book: "If $T : \mathbb{R}^{|X|} \to \mathbb{R}^{|X|}$ is a contraction mapping with modulus $\alpha \in (0, 1)$, then there exists a unique $J^* \in \mathbb{R}^{|X|}$ such that $J^* = TJ^*$.")

Similarly, for any stationary policy $\mu : \mathcal{X} \to \mathcal{U}$, the cost associated with μ is the unique solution to the following equation:

$$J_\mu(x) = \mathbb{E}\left[g(x, \mu(x), w) + \alpha J_\mu(f(x, \mu(x), w))\right]$$

That is, the only solution to the system of $n = |\mathcal{X}|$ equations $TJ = J$ is optimal.

3. **Optimal control.** A stationary policy μ^* is optimal if and only if $\mu^*(x)$ attains the minimum on the right hand side of the Bellman equation (1.53).

$$T_{\mu^*} J^* = TJ^* \iff J_{\mu^*} = J^*.$$

Proof of convergence of DP algorithm. Want to show: consider an arbitrary J. Consider the sequence $\{T^k J; k = 0, 1, \ldots\}$; want to show is Cauchy. For any $m > 0$, want to show

$$\|J_{m+k} - J_k\|_\infty < \epsilon.$$

We have

$$\begin{aligned}
\|J_{m+k} - J_k\|_\infty &= \left\| \sum_{i=1}^{m} (J_{k+i} - J_{k+i-1}) \right\|_\infty \\
&\leq \sum_{i=1}^{m} \|J_{k+i} - J_{k+i-1}\|_\infty \\
&\leq \sum_{i=1}^{m} \alpha^{k+i-1} \|J_1 - J_0\|_\infty \qquad \text{(by Theorem 1.22.6.3)} \\
&= \|J_1 - J_0\|_\infty \alpha^k \sum_{i=1}^{m} \alpha^{i-1} \\
&\leq \frac{\|J_1 - J_0\|_\infty \alpha^k}{1 - \alpha}
\end{aligned}$$

since by Theorem 1.22.6.3

$$\begin{aligned}
\|J_{\ell+1} - J_\ell\|_\infty &= \|TJ_\ell - TJ_{\ell-1}\|_\infty \\
&\leq \alpha \|J_\ell - J_{\ell-1}\|_\infty \\
&\leq \alpha^2 \|J_{\ell-1} - J_{\ell-2}\|_\infty \\
&\vdots \\
&\leq \alpha^\ell \|J_1 - J_0\|_\infty
\end{aligned}$$

Therefore $\{T^k J : k \geq 0\}$ is Cauchy, and complete. So $\lim_{h\to\infty} T^k J$ is well-defined.

□

By definition of $T^k J$ and the MDP assumption, the limit is the optimal cost $J^* = \lim_{k\to\infty} T^k J$.

Proof of statement about Bellman's Equation. Want to prove: J^* is the unique solution to the Bellman equation (1.53). First we will show it is a solution. For any $J : \mathcal{X} \to \mathbb{R}$,

$$\lim_{h\to\infty} T^k J = J^*.$$

So it must be that $TJ^* = J^*$ or else convergence is violated. Therefore J^* is a solution to the Bellman equation.

Suppose there exists $J' \neq J^*$ such that $TJ' = J'$. Then $T^3 J' = T^2 J' = TJ' = J'$, $T^k J' = J'$ for all k. then **???**

□

Proof of statement about Optimal control. Suppose that $TJ^* = T_\mu J^*$. Want to show that $J^* = J_\mu$.

$$J^* = TJ^* \text{ (by convergence of DP algorithm)} = T_\mu J^* \text{ (by hypothesis)}$$

So J^* is a fixed point of T_μ. By (2) T_μ has a unique fixed point given by J_μ. Therefore $J^* = J_\mu$.

Now suppose $J^* = J_\mu$. Want to show: $TJ^* = T_\mu J^*$.

$$J^* = J_\mu \iff T_\mu J^* = T_\mu J_\mu = J_\mu \text{ (by 2)} = J^* \text{ (by hypothesis)} = TJ^* \text{ (by 1)}$$

□

Optimality control using matrix notation: Let $X = \{1, \dots, n\}$. $p_{ij}(u) = \mathbb{P}(X_{k+1} = j \mid X_k = i, u_k = u)$.

$$\begin{aligned}
(TJ)(i) &= \min_{u\in\mathcal{U}(i)} \mathbb{E}\left[g(i,u,w) + \alpha J(f(i,u,w))\right] \\
&= \min_{u\in\mathcal{U}(i)} \sum_{j=i}^{n} P_{ij}(u)\left[g(i,u,j) + \alpha J(j)\right] \\
&= \min_{u\in\mathcal{U}(i)} \sum_{j=i}^{n} P_{ij}(u) g(i,u,j) + \alpha \sum_{j=i}^{n} P_{ij}(u) J(j) \\
\implies TJ(i) &= \min_{u\in\mathcal{U}(i)} \overline{g}(i,u) + \alpha \left[\underbrace{P(u)}_{n\times n \text{ matrix}} \underbrace{J}_{n\times 1 \text{ vector}} \right]_i \\
\iff \underbrace{TJ}_{n\times 1 \text{ vector}} &= \underbrace{\left[\min_{u\in\mathcal{U}(i)} \overline{g}(i,u)\right]}_{n\times 1 \text{ vector}} + \alpha \left[\underbrace{P(u)}_{n\times n \text{ matrix}} \underbrace{J}_{n\times 1 \text{ vector}} \right]
\end{aligned}$$

Question; given a stationary policy μ, how do I compute $J\mu$?

$$(T_\mu J)(i) = \overline{g}(i, \mu(i)) + \alpha J(j) \sum_{j=i}^{n} P_{ij}(\mu(i)) J(j)$$

Denote J as an n-vector, likewise with TJ. Denote

$$P_\mu = \begin{pmatrix} P_{11}(\mu(1)) & P_{12}(\mu(1)) & \cdots & P_{1n}(\mu(1)) \\ P_{21}(\mu(2)) & P_{22}(\mu(2)) & \cdots & P_{2n}(\mu(2)) \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1}(\mu(n)) & P_{n2}(\mu(n)) & \cdots & P_{nn}(\mu(n)) \end{pmatrix}$$

where for example $P_{12}(\mu(1))$ is the probability that if i have policy μ and I'm in state 1 I will transition to state 2. And

$$\overline{g}_\mu = \begin{pmatrix} \overline{g}(1, \mu(1)) \\ \vdots \\ \overline{g}(n, \mu(n)) \end{pmatrix}$$

Then $T_\mu J = \overline{g}_\mu + \alpha P_\mu J$. J_μ is the unique fixed point of T_μ.

$$\begin{aligned} T_\mu J_\mu &= J_\mu \\ J_\mu &= \overline{g}_\mu + \alpha P_\mu J_\mu \\ J_\mu &= (I - \alpha P_\mu)^{-1} g_\mu \end{aligned}$$

although this is a bad way to compute because if α is large (e.g. 0.9999), matrix could be very ill-conditioned since largest eigenvalue of a transition matrix is 1.

1.22.7 Value Iterations and Policy Iterations (Sections 2.2 and 2.3 of Bertsekas [2012b]; starts on p. 210 of pdf for Volume 3)

Question: how do I compute J^* and the optimal policy?

1. **VI: value iteration** (also called successive approximation)

   Pick k_0 and start with some J_0, apply the DP operator T k_0 times, output $T^{k_0} J_0 = \hat{J}$, use $\hat{J}$ as an approximation for J^*, find a policy ("greedy policy") μ such that $T\hat{J} = T_\mu \hat{J}$. Recall

   $$(T\hat{J})(x) = \min_{u \in \mathcal{U}(x)} \left\{ \mathbb{E}\left[g(x, u, w) + \alpha \hat{J}(f(x, u, w)) \right] \right\}$$

   Questions:

   (a) how should we pick k? (need error bound.)

(b) What is the difference between J_μ and J^*? Want to show: if $\hat{J}$ is close to J^*, then J_μ is also close to J^*. (cost of your policy in practice will be close to optimal.) Important distinction: $\hat{J} \neq J_{\mu_k}$ ($\hat{J}$ may not be achievable by any policy, let alone a stationary policy).

First question: trivial error bound: recall

$$\|T^k J - J^*\|_\infty = \|T^k J - T^k J^*\|_\infty \leq \alpha^k \|J - J^*\|_\infty$$

(see also Equation (7.9) in Bertsekas 3rd edition Vol. 1, p. 215 of pdf)

Theorem 1.22.7.1 (**Similar to Proposition 2.2.1 in Bertsekas [2012b]; in equation 7.17 (p. 216 of pdf) and 7.23 (p. 219 of pdf) in Bertsekas 3rd edition Vol. 1**). Monotonic error bound: for any vector J and state i with $k \geq 0$,

$$(T^k J)(i) + \underline{c}_k \leq (T^{k+1} J)(i) + \underline{c}_{k+1} \leq J^*(i) \leq (T^{k+1} J)(i) + \overline{c}_{k+1} \leq (T^k J)(i) + \overline{c}_k$$

where

$$\underline{c}_k = \frac{\alpha}{1-\alpha} \min_{i \in [n]} \left\{ (T^k J)(i) - (T^{k-1} J)(i) \right\}, \qquad \overline{c}_k = \frac{\alpha}{1-\alpha} \max_{i \in [n]} \left\{ (T^k J)(i) - (T^{k-1} J)(i) \right\}$$

So when gap between $\underline{c}_k$ and $\overline{c}_k$ is small, you know you are close to converging (because as k increases we know that value converges, so eventually the difference over which you are maximizing and minimizing goes to 0.)

Proof. We will only prove the lower bound (the proof of the upper bound is analogous). Let $\underline{\gamma}_1 := \min_{i \in [n]} \{(TJ)(i) - J(i)\}$. Then if $\boldsymbol{e}$ is the selection vector for the index of this minimum,

$$\begin{aligned} & \underline{\gamma}_1 \boldsymbol{e} \preceq TJ - J \\ \iff & J + \underline{\gamma}_1 \boldsymbol{e} \preceq TJ \end{aligned}$$

where $\preceq$ denotes an element-wise vector inequality. Next recall from (1.52) that $TJ \preceq \alpha J$, so $T\underline{\gamma}_1 \boldsymbol{e} \preceq \alpha \underline{\gamma}_1 \boldsymbol{e}$. Therefore we have from the above that $TJ + \alpha \underline{\gamma}_1 \boldsymbol{e} \preceq TJ + T\underline{\gamma}_1 \boldsymbol{e} \preceq TJ$. It also follows trivially from the above that $J + \underline{\gamma}_1 \boldsymbol{e} + \alpha \underline{\gamma}_1 \boldsymbol{e} \preceq TJ + \alpha \underline{\gamma}_1 \boldsymbol{e}$, so we have

$$\begin{aligned} & \iff J + (1+\alpha)\underline{\gamma}_1 \boldsymbol{e} \preceq TJ + \alpha \underline{\gamma}_1 \boldsymbol{e} \preceq T^2 J \\ \iff & J + (1+\alpha+\alpha^2)\underline{\gamma}_1 \boldsymbol{e} \preceq TJ + \alpha(1+\alpha)\underline{\gamma}_1 \boldsymbol{e} \preceq T^2 J + \alpha^2 \underline{\gamma}_1 \boldsymbol{e} \preceq T^3 J \end{aligned}$$

Applying this repeatedly yields

$$\begin{aligned} J + \left(\sum_{i=0}^{k} \alpha^i \right) \underline{\gamma}_1 \boldsymbol{e} & \preceq TJ + \left(\sum_{i=1}^{k} \alpha^i \right) \underline{\gamma}_1 e \\ & \preceq T^2 J + \left(\sum_{i=2}^{k} \alpha^i \right) \underline{\gamma}_1 e \\ & \vdots \\ & \preceq T^{k+1} J \end{aligned}$$

Take the limit as $k \to \infty$:

$$J + \underline{\gamma}_1 \frac{\boldsymbol{e}}{1-\alpha} \preceq TJ + \frac{\alpha}{1-\alpha}\underline{\gamma}_1 \boldsymbol{e} \preceq T^2 J + \frac{\alpha^2}{1-\alpha}\underline{\gamma}_1 \boldsymbol{e} \preceq \ldots \preceq T^k J + \frac{\alpha^k \boldsymbol{e}}{1-\alpha} \preceq \ldots \preceq J^*.$$

Recall:

$$\underline{c}_1 = \frac{\alpha}{1-\alpha} \min_{i \in [n]} \{(TJ)(i) - J(i)\} = \frac{\alpha}{1-\alpha}\underline{\gamma}_1$$

$$\implies J + \frac{\underline{c}_1}{\alpha}\boldsymbol{e} \preceq TJ + \underline{c}_1 \boldsymbol{e} \preceq T^2 J + \alpha \underline{c}_1 \boldsymbol{e} \preceq \ldots \preceq T^k J + \alpha^{k-1}\underline{c}_1 \boldsymbol{e} \preceq \ldots \preceq J^*, \tag{1.54}$$

Continuing further, since J is arbitrary, take $J = T^k J$. Then re-do analysis with $J = T^k J$. Then using

$$\underline{\gamma}_{k+1} = \min(T^{k+1}J)(i) - (T^k J)(i) = \underline{c}_{k+1}\left(\frac{1-\alpha}{\alpha}\right)$$

and the same argument will yield

$$T^k J + \frac{\underline{c}_{k+1}}{\alpha}\boldsymbol{e} \preceq T^{k+1}J + \underline{c}_{k+1}\boldsymbol{e} \preceq J^*. \tag{1.55}$$

Next, we will show that

$$\alpha \min_{i \in [n]}\{(TJ)(i) - J(i)\} \;\leq\; \min_{i \in [n]}(T^2 J)(i) - (TJ)(i)$$

Note that, by definition of $\underline{c}_1$ and $\underline{c}_2$, proving the above inequality is equivalent to showing that $\alpha \underline{c}_1 \leq \underline{c}_2$. We have from (1.54) that

$$\begin{aligned}
TJ + \underline{c}_1 \boldsymbol{e} \preceq T^2 J + \alpha \underline{c}_1 \boldsymbol{e} \qquad &\iff \qquad (1-\alpha)\underline{c}_1 \boldsymbol{e} \;\preceq\; T^2 J - TJ \\
&\iff \qquad (1-\alpha)\underline{c}_1 \;\leq\; \min_{i=1,\ldots,n}(T^2 J)(i) - (TJ)(i) \\
&\iff \qquad (1-\alpha)\underline{c}_1 \;\leq\; \frac{1-\alpha}{\alpha}\underline{c}_2 \\
&\iff \qquad \alpha \underline{c}_1 \;\leq\; \underline{c}_2
\end{aligned}$$

Then since from (1.54) $TJ + \underline{c}_1 \boldsymbol{e} \preceq T^2 J + \alpha \underline{c}_1 \boldsymbol{e}$ and $\alpha \underline{c}_1 \leq \underline{c}_2$, it follows that $TJ + \underline{c}_1 \boldsymbol{e} \;\leq\; T^2 J + \underline{c}_2 \boldsymbol{e} \;\leq\; J^*$. By a similar argument generalizing from (1.55), more generally for any vector J and state i with $k \geq 0$

$$(T^k J)(i) + \underline{c}_k \leq (T^{k+1}J)(i) + \underline{c}_{k+1} \leq J^*(i).$$

□

Theorem 1.22.7.2. Given $\tilde{J}$ we consider a "greedy policy" μ with respect to $\tilde{J}$, i.e. $T_\mu \tilde{J} = T\tilde{J}$. Then,

$$\|J_\mu - J^*\| \leq \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|.$$

Remark 10. Due to the MDP assumptions, there are only finitely many policies. So if we make the gap $\|\tilde{J} - J^*\|$ sufficiently small, we can eventually find the optimal policy.

Proof. We know that $T_\mu^k J^* \xrightarrow{k\to\infty} J_\mu$. So for all k

$$\|T_\mu^k J^* - J^*\| = \left\|\sum_{\ell=1}^{k} T_\mu^\ell J^* - T_\mu^{\ell-1} J^*\right\| \leq \sum_{\ell=1}^{k} \left\|T_\mu^\ell J^* - T_\mu^{\ell-1} J^*\right\| \leq \sum_{\ell=1}^{k} \alpha^{k-1} \|T_\mu l J^* - J^*\|$$
$$\leq \frac{\|T_\mu J^* - J^*\|}{1-\alpha}$$

Take limit as $k \to \infty$ (of the left side),

$$\|J_\mu - J^*\| \leq \frac{\|T_\mu J^* - J^*\|}{1-\alpha}$$

Now use the fact that μ is a greedy policy.

$$\begin{aligned}\|T_\mu J^* - J^*\| &\leq \|T_\mu J^* - T_\mu \tilde{J}\| + \|T_\mu \tilde{J} - J^*\| \\ &= \|T_\mu J^* - T_\mu \tilde{J}\| + \|T_\mu \tilde{J} - T J^*\| \\ &\leq 2\alpha \|J^* - \tilde{J}\|\end{aligned}$$

□

2. **Policy iteration (PI):** idea: start with an initial policy μ, compute J_μ, find an improving policy by acting greedily with respect to J_μ. Description: start with μ^0. For $k \in \{0, 1, \ldots\}$, given a policy μ^k, do two things:

   (a) Policy evaluation: compute J_{μ^k}. (Note: J_{μ^k} is the unique fixed point.)

   $$J_{\mu^k} = (I - \alpha P_{\mu^k})^{-1} \overline{g}_{\mu^k}.$$

   (b) Policy improvement: a new policy μ^{k+1} is obtained by acting greedily with respect to J_{μ^k}; that is, find μ^{k+1} solving

   $$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}.$$

   Note: if $T J_{\mu^k} = T_{\mu^k}$, we're done (because we've found the optimal policy).

Theorem 1.22.7.3 (Similar to Proposition 7.2.2 in 3rd Edition Vol. 1 of Bertsekas, p. 217 of pdf). If $T_{\overline{\mu}} J_\mu = T J_\mu$, then $J_{\overline{\mu}}(i) \leq J_\mu(i)$ for all states i and the inequality is strict for some i if μ is not optimal.

Proof. Since $T_\mu J_\mu = J_\mu$, we have for all states i

$$\begin{aligned}J_\mu(i) &= \overline{g}(i, \mu(i)) + \sum_{j=1}^{n} P_{ij}(\mu(i)) J_\mu(i) \\ &\geq \min_{u \in \mathcal{A}(i)} \overline{g}(i, u) + \sum_{j=1}^{n} P_{ij}(u) J_\mu(j) \\ &= (T J_\mu)(i) \\ &= (T_{\overline{\mu}} J_\mu)(i)\end{aligned}$$

So $J_\mu \geq T_{\overline{\mu}} J_\mu$. Apply $T_{\overline{\mu}}$ again,

$$J_\mu \geq T_{\overline{\mu}} J_\mu \geq T_{\overline{\mu}}^2 J_\mu \geq T_{\overline{\mu}}^3 J_\mu \geq \ldots \geq J_{\overline{\mu}}$$

So $J_\mu \geq J_{\overline{\mu}}$. Now we want to show that if μ is not optimal that the inequality is strict for some i; that is, if $J_\mu = J_{\overline{\mu}}$ then μ is optimal. Note that if $J_\mu = J_{\overline{\mu}}$

$$J_\mu = T_{\overline{\mu}} J_{\overline{\mu}} = T_{\overline{\mu}} J_\mu = \text{ (by assumption of Theorem) } TJ_\mu,$$

so J_μ is a fixed point of T. Since the only fixed point of T is the optimal one, μ is optimal.

□

Example 1.22.8. $X = \{1, 2\}, \mathcal{U} = \{u^1, u^2\}$. We have

$$P(u^1) = \begin{pmatrix} P_{11}(u^1) & P_{12}(u^1) \\ P_{21}(u^1) & P_{22}(u^1) \end{pmatrix} = \begin{pmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{pmatrix}, \qquad P(u^2) = \begin{pmatrix} P_{11}(u^2) & P_{12}(u^2) \\ P_{21}(u^2) & P_{22}(u^2) \end{pmatrix} = \begin{pmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{pmatrix},$$

Let $g(1, u^1) = 2, g(1, u^2) = 0.5, g(2, u^1) = 1, g(2, u^2) = 3$, and let $\alpha = 0.9$. Initialization: $\mu^0(1) = u^1, \mu^0(2) = u^2$. Policy evaluation: compute J_{μ^0} as the solution to the equation $T_{\mu^0} J_{\mu^0} = J_{\mu^0}$.

$$J_{\mu^0}(1) = 2 + 0.9 \left[\frac{3}{4} J_{\mu^0}(1) + \frac{1}{4} J_{\mu^0}(2) \right], J_{\mu^0}(2) \qquad = 3 + 0.9 \left[\frac{1}{4} J_{\mu^0}(1) + \frac{1}{4} J_{\mu^0}(2) \right]$$

System of two equations, two variables. Solving gives $J_{\mu^0}(1) = 24.12, J_{\mu^0}(2) = 25.96$.

Next, policy improvement. Find a policy $\mu^1 = (\mu^1(1), \mu^1(2))$ such that $T_{\mu^1} J_{\mu^0} = TJ_{\mu^0}$. So

$$\begin{aligned} (TJ_{\mu^0})(1) &= \min \left\{ 2 + 0.9 \left[\frac{3}{4} \cdot 24.12 + \frac{1}{4} \cdot 25.96 \right], 0.5 + 0.9 \left[\frac{1}{4} \cdot 24.12 + \frac{3}{4} \cdot 25.96 \right] \right\} \\ &= \min\{24.122, 23.45\} = 24.122 \end{aligned}$$

So $\mu^1(1) = u^2$. By a similar calculation,

$$(TJ_{\mu^0})(2) = \min\{23.12, 25.95\} = 23.12$$

So $\mu^1(2) = 1$. Next, compute

$$J_{\mu^1} = (I - \alpha P_{\mu^1})^{-1} \cdot g_{\mu^1} = \begin{pmatrix} 1 - \alpha\frac{1}{4} & -\alpha\frac{3}{4} \\ -\alpha\frac{3}{4} & 1 - \alpha\frac{1}{4} \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} = \begin{pmatrix} 7.31 \\ 7.67 \end{pmatrix}$$

If you do another iteration, you will find $T_{\mu^2} J_{\mu^1} = TJ_{\mu^1}$, so $\mu^2 = \mu^1$ and you're done.

Now, want to compare value iteration vs. policy iteration. Assume that we have 1 state. Each policy maps the state to an action, but there is only one state, so the policy is just an action. In particular, it defines a line $g_\mu + \alpha P_\mu J$. Then

$$TJ = \min_u \{g_\mu + \alpha P_\mu J\}$$

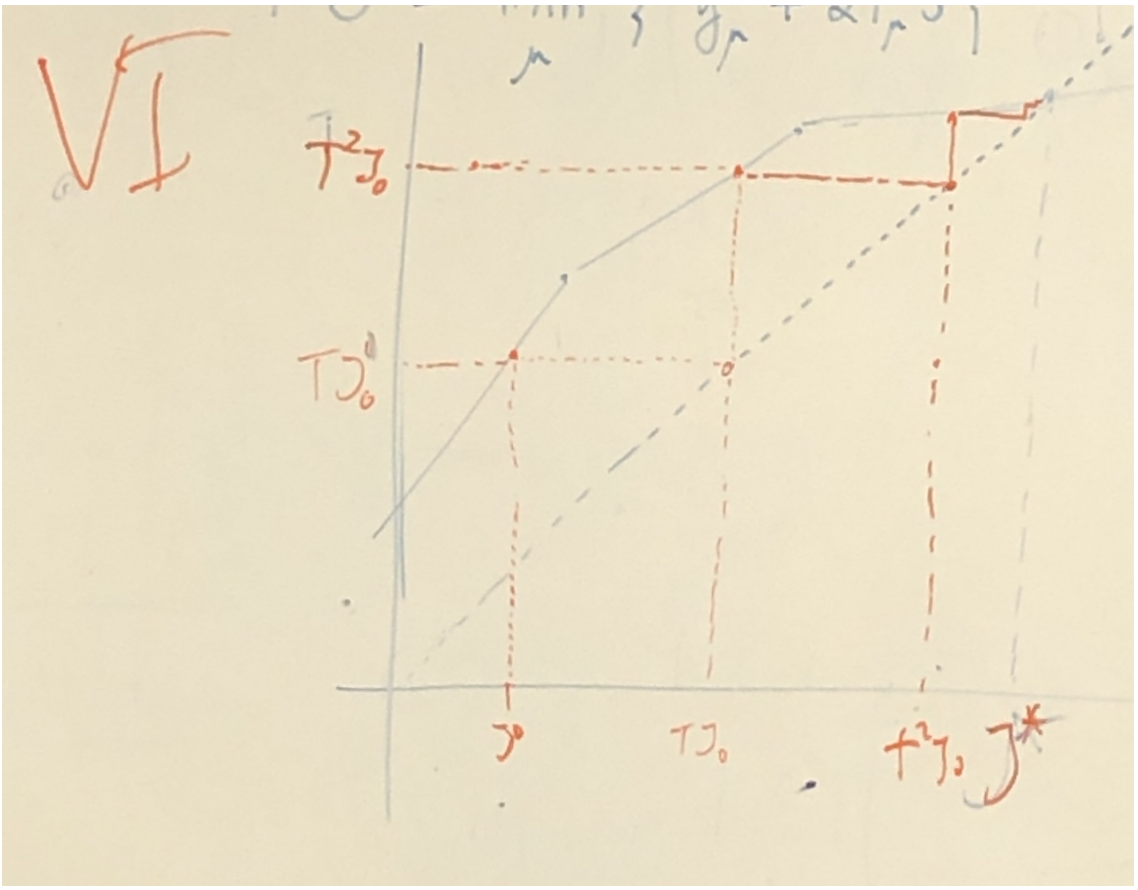

Figure 1.15: Iteration of VI method. Start with J_0, get TJ_0, plug in to T, etc. Note that the optimal solution is the intersection of the TJ function and the 45 degree line, since the optimal policy satisfies $TJ^* = J^*$.

The function is increasing, concave, and piecewise linear in J. Each piece corresponds to a policy. Value iteration: attempts to approximate J^*. Never reaches J^*, but since you have finitely many polices, you eventually reach the optimal policy. See Figure 1.15.

Policy iteration is different. You fall into a policy, representing one piece of the function (one line segment—each one corresponds to a policy). Then compute J_{μ^1}, a fixed point of T_{μ^1}. T_{μ^1} is the extension of the line segment corresponding to policy μ^1. So J_{μ^1} is the point where that line intersects the 45 degree line $T_{\mu^1} J = J$. Then you plug that point in and solve again, finding the next policy. Generally this approach is way faster (although in the worst case it can be very slow). See Figure 1.16.

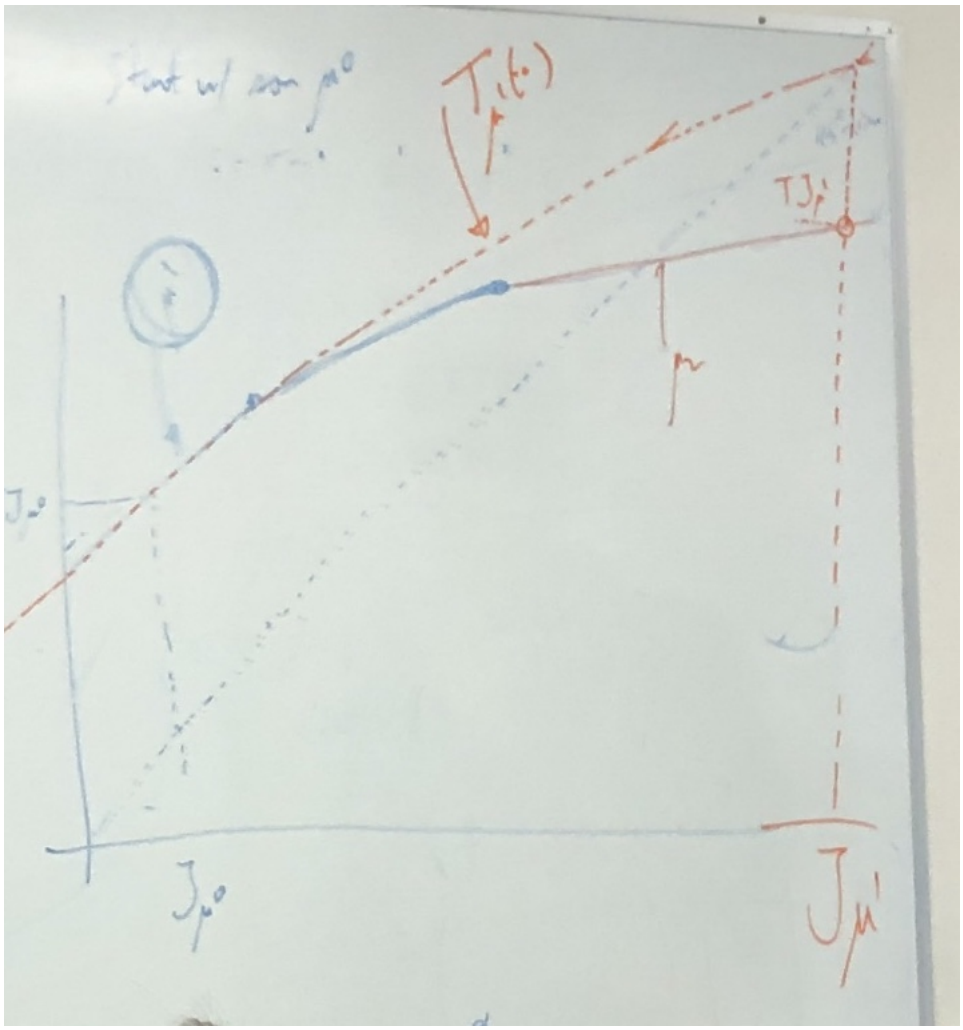

Figure 1.16: Iteration of PI method.

3. **Optimistic policy iteration (blend of PI and VI; section 2.3.3 of Bertsekas [2012b])** Start with an initial J_0. Let $\{m_k : k = 0, 1, \ldots\}$ be any sequence of positive integers (say $\{1, 1, 1, \ldots\}$. Do the following:

   (a) Take a greedy decision with respect to J_k to get the policy μ^k: $T_{\mu^k} J_k = T J_k$. Then get $J_{k+1} = T_{\mu^k}^{m_k} J_k$. If $m_k = 1$ for all k, we get value iteration: $J_{k+1} = T_{\mu^k} J_k = T J_k$ for all k. If $m_k = \infty$

for all k (or is sufficiently large), we get $J_{k+1} = J_{\mu^k}$, so $T_{\mu^{k+1}} J_{k+1} = T J_{k+1}$. This is policy iteration. So this method subsumes PI and VI, and it doesn't require inverting a matrix like PI. Also we have lots of flexibility; converges to the optimal policy for any sequence (will prove next time).

Theorem 1.22.7.4 (Proposition 2.3.2 of Bertsekas [2012b]). *J_k coverges to J^* and μ^k is optimal for all k sufficiently large.*

Proof. "the same sequence of policies will be obtained" (p. 105): true because the costs differ only by constants, so the policy resulting in the optimum for both cost functions will be the same.

Prove by induction on k that $\overline{J}_k \leq T\overline{J}_{k-1}$ and $T\overline{J}_k \leq \overline{J}_k$.

Base case: ($k = 1$) we want to show that

$$\overline{J}_1 \leq T\overline{J}_0, \qquad T\overline{J}_0 \leq \overline{J}_0.$$

By definition, $T_{\mu^0}\overline{J}_0 = T\overline{J}_0 \leq \overline{J}_0$ where the first step follows by definition of μ^0 and the second follows by our construction of $\overline{J}_0$. So then $T^m_{\mu^0}\overline{J}_0 \leq T^{m-1}_{\mu^0}\overline{J}_0$ for all m.

By definition (and using $\overline{J}_1 = T^{m_0}_{\mu^0}\overline{J}_0$)

$$T_{\mu_1}\overline{J}_1 = T\overline{J}_1 \leq T_{\mu^0}\overline{J}_1 = T^{m_0+1}_{\mu^0}\overline{J}_0 \leq T^{m_0}_{\mu^0}\overline{J}_0 = \overline{J}_1$$

where the second-to-last step follows from $T^m_{\mu^0}\overline{J}_0 \leq T^{m-1}_{\mu_0}\overline{J}_0$ for all m. But then $\overline{J}_1 \leq T_{\mu^0}\overline{J}_0$, because to get $\overline{J}_1$ we applied T_{μ^0} $m_0 \geq 1$ times. Therefore $\overline{J}_1 \leq T\overline{J}_0$.

Inductive step: Assume that $\overline{J}_k \leq T\overline{J}_{k-1}$ and $T\overline{J}_k \leq \overline{J}_k$. We want to show that $\overline{J}_{k+1} \leq T\overline{J}_k$ and $T\overline{J}_{k+1} \leq \overline{J}_{k+1}$.

By definition, $T_{\mu^k}\overline{J}_k = T\overline{J}_k \leq \overline{J}_k$ where the last step follows by the inductive step. Therefore $T^m_{\mu^k}\overline{J}_k \leq T^{m-1}_{\mu^k}\overline{J}_k$ for all m (same argument as above).

Finally,

$$T_{\mu^{k+1}}\overline{J}_{k+1} = T\overline{J}_{k+1} \leq T_{\mu^k}\overline{J}_{k+1} = (\text{by definition of } \overline{J}_{k+1}) T^{m_k+1}_{\mu^k}\overline{J}_k \leq (\text{by above result}) \; T^{m_k}_{\mu^k}\overline{J}_k$$
$$= \overline{J}_{k+1} \leq T_{\mu^k}\overline{J}_k = T\overline{J}_k.$$

This gives us that $\overline{J}_{k+1} \leq T\overline{J}_k$ and $T\overline{J}_{k+1} \leq \overline{J}_{k+1}$. Now, note that

$$\begin{aligned} & \overline{J}_1 \leq T\overline{J}_0 \\ \iff \quad & \overline{J}_2 \leq T\overline{J}_1 \leq T^2\overline{J}_0 \\ \iff \quad & \overline{J}_3 \leq T\overline{J}_2 \leq T^2\overline{J}_1 \leq T^3\overline{J}_0 \end{aligned}$$

which leads to $\overline{J}_k \leq T^k\overline{J}_0$. Then to show $J^* \leq \overline{J}_k$ for all k, note that

$$T\overline{J}_0 \leq \overline{J}_0 \implies J^* \leq \overline{J}_0$$

because $J^= \lim_{k\to\infty} T^k\overline{J}_0$. Also for all ℓ

$$T^\ell\overline{J}_k \leq T^{\ell-1}\overline{J}_k \leq \ldots \leq T\overline{J}_k \leq \overline{J}_k.$$

□

Definition 1.22.8 (**One-step look-ahead policy; section 2.3.4, p. 106 of Bertsekas [2012b]**). If $\tilde{J}$ is an approximation to J^*, then a one-step look-ahead policy based on $\tilde{J}$ is a policy $\overline{\mu}$ such that $T_{\overline{\mu}}\tilde{J} = T\tilde{J}$.

Note: can be an arbitrary vector $\tilde{J}$. Also if $\tilde{J} = J_\mu$ for some policy μ, then $J_{\overline{\mu}} \leq J_\mu$. Two-step lookahead:

$$T_{\overline{\mu}}(T\tilde{J}) = T(T\tilde{J}).$$

Theorem 1.22.7.5 (**Proposition 2.3.3, p. 107 of Bertsekas [2012b]**). Let $\overline{\mu}$ be the one-step look-ahead policy based on $\tilde{J}$, i.e., $T_{\overline{\mu}}\tilde{J} = T\tilde{J}$. Let $\hat{J} := T_{\overline{\mu}}\tilde{J} = T\tilde{J}$. Then

(a) If $\hat{J} \leq \tilde{J}$, then $J_{\overline{\mu}} \leq \hat{J}$.

(b) $\|J_{\overline{\mu}} - \hat{J}\| \leq \frac{\alpha}{1-\alpha}\|\hat{J} - \tilde{J}\|$, $J_{\overline{\mu}} - J^*\| \leq 2\alpha/(1-\alpha)\|\tilde{J} - J^*\|$, and $\|J_{\overline{\mu}} - J^*\| \leq 2/(1-\alpha)\|\tilde{J} - \hat{J}\|$.

Example 1.22.9 (**Example 2.3.1, p. 109 of Bertsekas [2012b]**). Note that

$$T\tilde{J}(1) = \min\{0 + \alpha\tilde{J}(2), 2\alpha\epsilon + \alpha\tilde{J}(1)\} = \min\{\alpha\epsilon, \alpha\epsilon\} = \alpha\epsilon$$

and $T\tilde{J}(2) = 0 + \alpha\tilde{J}(2) = \alpha\epsilon$.

So, with the one-step lookahead

$$J_{\overline{\mu}}(1) = \frac{2\alpha\epsilon}{1-\alpha} = \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|$$

So

$$\|J_{\overline{\mu}} - J^*\| = \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|$$

so the bound is tight.

Proposition 1.22.7.6 (**Proposition 2.3.4, p. 111 of Bertsekas [2012b]**).

4. **Linear programming (LP)**

   Basic form: $\max c^T x$ s.t. $Ax \leq b, x \geq 0$. Equivalent dual: $\min \pi^T b$ s.t. $\pi^T A \geq c^T, \pi \geq 0$. Dual problem may have a smaller constraint set (number of variables in dual equals number of constraints in the primal). In primal problem: number of constraints equals number of state-action pairs, number of variables equals number of states. So both number of constraints and variables is large. So dual doesn't generally help a lot. But it turns out that for a variety of important problems there are ways to exploit the LP structure.

exam: inventory, revenue management, VI (variants, applications in weird contexts), PI question for sure. so at least those 4 questions. maybe a 5th on one of these topics. optimistic PI maybe. LP no.

1.22.8 Scheduling and Multiarmed Bandit Problems (Section 1.3 of Bertsekas [2012b])

We will examine the multiarmed bandit (MAB) in an infinite horizon discounted reward setting. (One of the few high-dimensional DPs that can be solved feasibly.) See setup on p. 22 of Bertsekas [2012b]. Other notation: $x_k^\ell \in S^\ell$ (state space).

$$J(x^1, \dots, x^n) = \max_{u \in [n]} \left\{ R^u(x^u) + \alpha \mathbb{E}[J(x^1, \dots, x^{u-1}, f^u(x^u, w^u), x^{u+1}, \dots, x^n) \right\}$$

(similar to equation (1.12) on p. 23, but without M retirement reward.)

Note this is high-dimensional: if each arm has 10 states, then the number of possible states is 10^m.

Key features:

1. States of idle projects remain fixed.
2. Rewards received only depend on the state of selected arms.
3. Only one product can be chosen.

Gittins in 1970s showed that an optimal policy for this problem is an index rule (equation (1.11) on p. 23). The index associated with each project can be computed by solving a 1 dimensional dynamic program. (i.e., solve n 1-dimensional problems rather than a single n-dimensional problem—much more feasible when n is large.)

First step to proving optimality of this approach: generalize the problem to allow the option of quitting at any time k. Key: the retirement reward M will be crucial in defining our index function.

Theorem 1.22.8.1 (**Proposition 1.3.4 in Bertsekas [2012b], p. 29**). For each project ℓ, there exists a function $m^\ell : S^\ell \to \mathbb{R}$ such that at each time k an optimal policy is given by the rule

- Retire if $M > \max_{h \in [n]} \{m^h(x_k^h)\}$,
- Work on project ℓ if $m^\ell(x_k^\ell) = \max_{h \in [n]} \{m^h(x_k^h)\} \geq M$,

where the index function $m^\ell(x^\ell)$ is as defined in (1.56).

We will get some preliminary results to be able to prove this.

Lemma 1.22.8.2 (**Proposition 1.3.1 in Bertsekas [2012b], p. 29**). Let

$$B := \max_\ell \max_{x^\ell} |R^\ell(x^\ell)|.$$

For fixed $x \in S_1 \times \ldots \times S_n$, the optimal reward function (as a function of M) $M \mapsto J(x, M)$ has the following properties:

(a) $M \mapsto J(x, M)$ is convex and monotonically nondecreasing.

(b) $J(x, M)$ is constant for $M \leq -B/(1-\alpha)$.

(c) $J(x, M) = M$ for all $M \geq B/(1-\alpha)$.

(See Figure 1.17.)

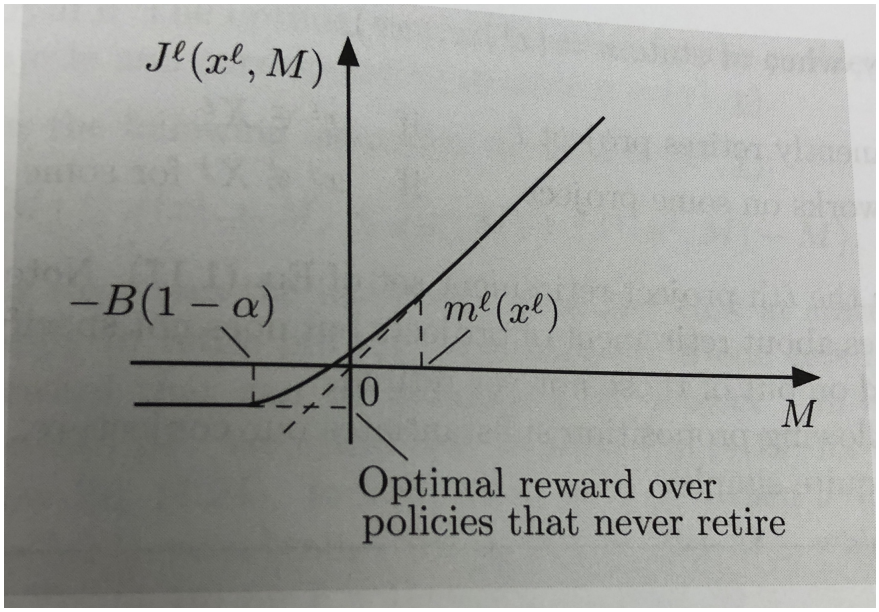

Figure 1.17: Figure 1.3.1 in Bertsekas [2012b], p. 25. Form of the ℓth project reward function $J^\ell(x^\ell, M)$ for fixed x^ℓ and definition of the index $m^\ell(x^\ell)$.

Proof. (a) inductive step: recall that maximum of convex functions is convex. for each u, get convex functions inside expectation. expectation preserves convexity, adding a constant preserves convexity, max preserves convexity.

(b) Note that the absolute worst reward you can ever get if you never retire is $-B/(1-\alpha)$ (sum of infinite series if you get $-B$ every time). So if M is worse than that, never makes sense to quit.

(c) Reverse of last argument.

□

Consider a situation with only one project (as on bottom of p. 24). There is a minimal value $m^\ell(x^\ell)$ of M for which $J^\ell(x^\ell, M) = M$ (the smallest M that would cause you to want to retire rather than work on your one project). That is,

$$m^\ell(x^\ell) = \inf\{Q : Q = J^\ell(x^\ell, Q)\}. \tag{1.56}$$

Definition 1.22.9 (Index function; p. 24 of Bertsekas [2012b]). The function $m^\ell(x^\ell)$ as defined in (1.56) is called the **index function** of project ℓ.

The index function is the retirement reward for which we are indifferent between retiring and operating the project when at state x^ℓ. Note that $m^\ell(x^\ell) \leq B/(1-\alpha)$, because per Lemma 1.22.8.2, we would always want to retire if $M > B/(1-\alpha)$.

Total number of curves upper bounded by number of states times number of projects (as opposed to number of states raised to the number of projects—so big reduction). Also, each curve corresponds to a one-dimensional DP.

How do you find $m^\ell(x^\ell)$ in practice? Easy to evaluate at a particular M; solve the DP (value iteration, policy iteration, etc.). Use a bisection algorithm to find it (similar to finding a zero of a function). Have a bounded interval to start with: $[-B/(1-\alpha), B/(1-\alpha)]$. For each DP, number of state-action pairs is number of states times 2 (only 2 actions: retire or continue).

Definition 1.22.10 (**Project-by-Project Retirement Policy (PPR; pp. 25-26 of Bertsekas [2012b])**). In a single project problem, the project is operated continuously up to the time that its state falls into the **retirement set**

$$X^\ell = \{x^\ell : m^\ell(x^\ell) < M\}. \tag{1.57}$$

A **project-by-project retirement policy** permanently retires projects in the same way as if they were the only project available:

- Permanently retire project ℓ if $x^\ell \in X^\ell$,
- work on some project if $x^j \notin X^j$ for some j (where X^j is the retirement set (1.57)).

Note that a PPR policy decides about retirement of projects but does not specify the project to be worked on out of those not yet retired.

Proposition 1.22.8.3 (**Proposition 1.3.2 in Bertsekas [2012b]**). There exists an optimal PPR policy.

Proof. How to get (1.20):

$$\begin{aligned} m^\ell(x^\ell) \geq M \quad \implies M &\leq R^\ell(x^\ell) + \alpha \mathbb{E}[J^\ell(f^\ell(x^\ell, w^\ell), M)] \\ &\leq R^\ell(x^\ell) + \alpha \mathbb{E}[J(x^1, \dots, x^{\ell-1}, f(x^\ell, w^\ell), x^{\ell+1}, \dots, x^n, M)] \\ &= L^\ell(x, M, J). \end{aligned}$$

□

Proposition 1.22.8.4 (**Proposition 1.3.3 in Bertsekas [2012b]**). For fixed x, let K_M denote the (random) retirement time under an optimal policy when the retirement reward is M. Then for all M for which $\partial J(x, M)/\partial M$ exists, we have

$$\frac{\partial J(x, M)}{\partial M} = \mathbb{E}\left[\alpha^{K_M} \mid x_0 = x\right].$$

Proof. pp. 28- 29, [Bertsekas, 2012b].

□

Let T_ℓ be the retirement time of project ℓ if it were the only project available and let T be the retirement time for the multiproject problem. We have $T = T_1 + T_2 + \ldots + T_n$ because the state of idle projects doesn't change. Further, all the T_ℓ are independent. Therefore

$$\mathbb{E}[\alpha^T] = \mathbb{E}[\alpha^{\sum_{\ell=1}^n T_\ell}] = \prod_{\ell=1}^n \mathbb{E}[\alpha^{T_\ell}].$$

Then Proposition 1.22.8.4 yields

$$\frac{\partial J(x,M)}{\partial M} = \prod_{\ell=1}^n \frac{\partial J^\ell(x^\ell, M)}{\partial M}. \tag{1.58}$$

We are now ready to prove Theorem 1.22.8.1.

Proof of Theorem 1.22.8.1. notes for end of proof: The function $M \mapsto J(x,M)$ and $M \mapsto L^\ell(x,M,J)$ coincide at $M = m(x)$. The derivatives of the two functions are the same for all $M \leq m(x)$. Therefore for all $M \leq m(x)$, $J(x,M) = L^\ell(x,M,J)$, so it's optimal to choose project ℓ.

□

1.22.9 Approximate DP: Q-Learning (Section 6.3.3 of Bertsekas [2012a], Sections 2.2.3, 2.5.3, and 6.1 - 6.6.1 of [Bertsekas, 2012b])

We'll focus on large-scale DP under a discounted cost criterion. Interested in setting when the number of states is extremely large (say 10^{100}).

How do we adapt VI for large-scale setting? $VI : TJ \to J$.

$$(TJ)(x) = \min_{u \in U(x)} \sum_{y \in S} P_{x,y}(u)[g(x,u,y) + \alpha J(y)].$$

How do we compute the sum when the state space is so large? A potential approach is to use Monte Carlo simultion. We have that $Y \sim P_{x,\cdot}(u)$, and the sum is $\mathbb{E}[g(x,u,Y) + \alpha J(Y)]$. Suppose we want to compute $\mathbb{E}[F(Z)]$ for some function F. The true value is $\sum_{z \in S} \mathbb{P}(z)F(z)$, but if $|S|$ is intractably large, consider an IID sample $z_1, \ldots, z_K$ drawn from Z. The Strong Law of Large Numbers tells us $\frac{1}{k}\sum_{k=1}^K F(z_k) \xrightarrow{a.s.} \mathbb{E}[F(Z)]$. If $K \ll |S|$, this sample is much easier to compute.

How large should the number of samples be? Recall Chebyshev's Inequality (**??**):

$$\mathbb{P}\left(\left|\frac{1}{K}\sum_{k=1}^K F(z_k) - \mathbb{E}[F(Z)] > \right| \epsilon\right) \leq \delta$$

if $K \geq \frac{1}{\delta}\frac{\text{Var}(F(Z))}{\epsilon^2}$.

We will combine Monte Carlo simulation with VI. To avoid the sum, for each $u \in U(x)$, we can sample $y_1, \ldots, y_K$ according to $P_{x,\cdot}(u)$ and approximate

$$\sum_{y \in S} P_{x,y}(u)[g(x,u,y) + \alpha J(y)] \approx \frac{1}{K}\sum_{k=1}^{K}\left[g(x,u,y_k) + \alpha J(y_k)\right].$$

Do we do this for each $u \in U(x)$? (Do we generate a new sample for each $u \in U(x)$?) What about for each x—do I have to update all states simultaneously? Can I update one state at a time?

Definition 1.22.11. Let

$$Q_k(x_k, u_k) := \mathbb{E}\left[g_k(x_k, u_k, y_k) + J_{k+1}(f_k(x_k, u_k, y_k))\right] = \sum_{y \in S} P_{x,y}(u)\left[g_k(x_k, u_k, y_k) + \alpha J^*(y)\right].$$

Let the optimal Q^* be defined by

$$Q^*(x,u) := \sum_{y \in S} P_{x,y}(u)\left[g(x,u,y) + \alpha J^*(y)\right], \qquad J^*(x) = \min_{u \in U(x)} Q^*(x,u).$$

(Note that this is simply a re-writing of Bellman's Equation. See Section 6.3.3 of Bertsekas [2012a], p.339 - 340.)

Suppose we're given $Q_k(\cdot,\cdot)$ based on sample $y_1, y_2, \ldots, y_{k-1}$. Then

$$Q_{k+1}(x,u) = \frac{1}{k}\sum_{\ell=1}^{k}[g(x,u,y_\ell) + \alpha \min_{v \in U(y_\ell)} Q_\ell(y_\ell, v)].$$

Note: for each iteration, we only have one new sample.

$$Q_{k+1}(x,u) = \frac{1}{k}\left[g(x,u,y_k) + \alpha \min_{v \in U(y_\ell)} Q_k(y_\ell, v)\right] + \frac{k-1}{k} \cdot \underbrace{\frac{1}{k-1}\sum_{\ell=1}^{k-1}[g(x,u,y_\ell) + \min_{v \in U(y_\ell)} Q_\ell(y_\ell, v)]}_{Q_k(x,u)}.$$

Therefore

$$\begin{aligned} Q_{k+1}(x,u) &= \frac{1}{k}\left[g(x,u,y_k) + \alpha \min_{v \in U(y_\ell)} Q_k(y_\ell, v)\right] + \underbrace{\frac{k-1}{k}}_{=1-1/k} Q_k(x,u) \\ &= Q_k(x,u) + \frac{1}{k}\left[g(x,u,y_k) + \alpha \min_{v \in U(y_\ell)} Q_k(y_\ell, v) - Q_k(x,u)\right]. \end{aligned}$$

Q-learning (algorithm from section 6.6.1 of Bertsekas [2012b], p. 496):

$$Q_{k+1}(x,u) = Q_k(x,u) + \gamma_k \left[g(x,u,y_k) + \alpha \min_{v \in U(y_\ell)} Q_k(y_\ell, v) - Q_k(x,u) \right].$$

where γ_k is the step size.

Theorem 1.22.9.1 (Section 6.6.1 of Bertsekas [2012b], p. 496). If each $Q(x,u)$ is updated infinitely often and $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$, then $Q_k \to Q^*$.

(See Section 6.3 of Bertsekas [2012b] for this material.) How do we deal with a huge number of states? Our goal is to compute $\{J^*(x) : x \in S\}$, but $|S|$ is huge. Approximate $J^*(x) \approx \sum_{k=1}^{K} \phi_k(x) \cdot r_k = (\Phi r)(x)$, where $\Phi \in \mathbb{R}^{|S| \times K}$ and $r \in \mathbb{R}^K$. Important: we do not store Φ explicitly. We compute $\phi_k(x)$ as needed.

Our goal is to find r such that Φr is close to J^*. We do this by projecting J^* onto the span of Φ; that is, we would like to compute

$$\min_{r \in \mathbb{R}^K} \sum_{x \in S} \pi(x) \left((\Phi r)(x) - J^*(x) \right)^2. \tag{1.59}$$

More compact notation:

$$\|\Phi r - J^*\|_\pi^2 := (\Phi r - J^*)^T D (\Phi r - J^*)$$

where

$$D = \begin{pmatrix} \pi(1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \pi(S) \end{pmatrix}.$$

and $\|\cdot\|_\pi$ is a weighted norm under the stationary probability distribution π of the transition matrix P. For any vector J

$$\|J\|_\pi = \sqrt{\sum_{x \in S} \pi(x) |J(x)|^2}.$$

Then we can write (1.59) as

$$\min_{r \in \mathbb{R}^K} \|\Phi r - J^*\|_\pi^2$$

Instead of summing over all states, we sample the states according to $\pi(\cdot)$, so $x_1, x_2, \dots, x_n$. Then we find

$$\min_{\pi \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^{n} \left((\Phi r)(x_i) - J^*(x_i) \right)^2.$$

However, this problem is still not tractable because we do not know J^*. Idea: leverage the "closed form" expression for the optimal r^* (if we have J^*)

$$\begin{aligned} 0 &= \nabla_r \|J^* - \Phi r\|_\pi^2 \\ &= \nabla_r \left[(\Phi r - J^*)^T D(\Phi r - J^*)\right] \\ &= 2\Phi^T D(\Phi r - J^*) \\ \implies \quad r^* &= (\Phi^T D\Phi)^{-1}\Phi^T D J^*, \\ \Phi r^* &= \Phi(\Phi^T D\Phi)^{-1}\Phi^T D J^*. \end{aligned}$$

Observe that $\Phi(\Phi^T D\Phi)^{-1}\Phi^T D$ is a projection matrix Π.

where Π is the projection onto the basis functions ϕ; projection with respect to $\|\cdot\|_\pi$, and π is the stationary distribution of the transition matrix P **(key: only one transition matrix, as in optimal stopping; see below)**.

Intuition: start with an initial J. Do approximate value iteration:

$$\Pi T J \in \mathbb{R}^{|S|}$$

project TJ into the K-dimensional vector space spanned by $\phi_1, \dots, \phi_K$. Don't keep track of the full TJ; instead, just the K coefficients.

Approximate VI: $(\Pi T)^k J$. Caveat: this update can lead to (??).

Example 1.22.10 (Example 6.4.2 in Bertsekas [2012b], p. 472)**.**

Proposed method: start at $r^{(0)} \in \mathbb{R}^K$. Iteration:

$$\begin{aligned} r^{(k+1)} &= \arg\min_{r \in \mathbb{R}^K} \|\Phi r - T\Phi r^{(k)}\|_\pi^2 \\ \iff \quad \Phi r^{(*k+1)} &= \Pi T(\Pi r(^{(k)}), \end{aligned}$$

where $\Pi = \Phi(\Phi^T D\Phi)^{-1}\Phi^T D$, with D as defined before.

Lemma 1.22.9.2 (**Lemma 6.3.1 in Bertsekas [2012b], p. 427**)**.** If $\Pi' P = \Pi'$, then $\|PJ\|_\Pi \le \|J\|_\Pi$.

Proof.

$$\|PJ\|_\Pi^2 = \sum_{x \in S} \pi(x) \left[(PJ)(x)\right]^2 = \sum_{x \in S} \pi(x) \left(\sum_{y \in S} P_{x,y} J(y)\right)^2 \le \text{ (by Jensen's inequality) } \sum_{x \in S} \pi(x) \sum_{y \in S} P_{x,y} (J(y))^2 = \sum_{y \in S} [J(y)]^2$$

□

Lemma 1.22.9.3 (Proposition 6.3.1(a) in Bertsekas [2012b], p. 429). Consider a policy μ. Let P be the transition matrix associated with μ. If $\Pi' P = \Pi'$ then $\|T_\mu J - T_\mu \overline{J}\|_\Pi \leq \alpha \|J - \overline{J}\|_\Pi$.

Proof.

$$\|T_\mu J - T_\mu \overline{J}\|_\Pi^2 = \sum_{x \in S} \pi(x) \left[(T_\mu J)(x) - (T_\mu \overline{J})(x)\right]^2 = \sum_{x \in S} \pi(x) \left[\alpha \sum_y P_{x,y} \left(J(y) - \overline{J}(y)\right)\right]^2 = \alpha^2 \sum_{x \in S} \pi(x) \left[(P(J - \overline{J}))(x)\right]^2 = \alpha$$

because

$$(T_\mu J)(x) = g(x, \mu(x)) + \alpha \sum_y P_{x,y}(\mu(x)) J(y).$$

□

Π is expansive under the sup norm. However...

Lemma 1.22.9.4 (Proposition 6.3.1(a) in Bertsekas [2012b], p. 429). The projection matrix Π is nonexpansive with respect to $\|\cdot\|_\Pi$; i.e., $\Pi J\|_\Pi \leq \|J\|_\Pi$.

Proof. We begin by observing that $J - \Pi J$ is orthogonal to ΠJ.

$$\Pi(J - \Pi J) = \Pi J - \Pi \Pi J = 0.$$

So $J - \Pi J$ is in the nullspace of Φr. Then

$$\|J\|_\pi = \|J - \Pi J\|_\pi + \|\Pi J\|_\pi \geq \|\Pi J\|_\pi.$$

□

We know have that ΠT_μ is a contraction mapping with respect to $\|\cdot\|_\pi$, where π is the stationary distribution with respect to P_μ. **This result ends up showing that $Q_{k+1} \to \hat{Q}$ because $\Pi \circ F$ is a contraction mapping under $\|\cdot\|_\pi$ and $\hat{Q}$ is the unique fixed point of $\Pi \circ F$; that is, $\Pi F \hat{Q} = \hat{Q}$.** How do we compute $(\Pi T)J$ with $J = \Phi r^{(k)}$?

1. Sample $x_1, \ldots, x_n$ from π.
2. Solve

$$r^{(k+1)} := \arg\min_{r \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^n \left[(\Phi r)(x_i) - (\Phi r^{(k)})(x_i)\right]^2$$

How do you get π? Simulate using a Markov chain for as long as needed; eventually converges to stationary distribution.

Note: divergence can occur if the wrong norm is used. We have convergence if

1.
$$(TJ)(x) = \sum_{y \in S} P_{x,y}(g(x,y) + \alpha J(y))$$
(only one action)

2. $\pi^T P = \pi^T$.

Heuristic for Approximate VI for general MDP. Initialization: given $\Phi = \begin{bmatrix} \phi_1 & \cdots & \phi_k \end{bmatrix}$ and $r^{(0)}$.

1. Given $r^{(k)}$, sample $x_1, \ldots, x_n$ from a distribution π. Then
$$r^{(k+1)} := \arg\min_{r \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^{n} \left[(\Phi r)(x_i) - (T\Phi r^{(k)})(x_i) \right]^2.$$
$$(T\Phi r)(x) = \min_{u \in U(x)} \sum_{y \in S} P_{x,y}(u) \left[g(x,u,y) + \alpha(\Phi r)(y) \right]^2$$
2.

(can't prove anything about this, just a "hack," but seems to work.)

1.22.10 Optimal Stopping (Section 6.6.4 of Bertsekas [2012b], p. 504)

Key insight: in an optimal stopping problem, there is one transition matrix.

Want a high-dimensional MC. ($X_t \in S : t = 1, 2, \ldots$)

DP operator:

$$(TJ)(x) = \max\{g(x), h(x) + \alpha \sum_{y \in S} P_{xy} J(y)\} = h(x) + \max\{g(x) - h(x), \alpha \sum_{y \in S} P_{x,y} J(y)\}.$$

$$TJ = \max\{g, \alpha P J\}.$$

Still suffer from curse of dimensionality even though we only have one DP.

$$Q^*(x) := \alpha \sum_{y \in S} \phi_{x,y} J^*(y) = \underbrace{\alpha \sum_{y \in S} P_{x,y} \max\{g(y), Q^*(y)\}}_{(FQ^*)(x)}, \qquad J^*(x) = \max\{ \underbrace{g(x)}_{\text{reward for stopping}}, \underbrace{Q^*(x)}_{\text{reward for continuing}} \}.$$

Q^* is the optimal value you get if you continue—the **optimal continuation function.**

$$Q^* = \alpha P J^* = \alpha P \max\{g, Q\} \qquad \iff \qquad Q^* = FQ^*$$

where $FQ = \alpha P \max\{g, Q\}$. So Q^* is the fixed point of a contraction mapping.

We'll approximate Q^* in general as a sequence of coefficients $r^{(1)}, r^{(2)}, \ldots$. Given $r^{(k)}$,

1. Sample $x_{k+1} \sim \pi$, where π is the stationary distribution of P. $y_{k+1} \sim P_{x_{k+1}}$.
2. $$r^{(k+1)} := \arg\min_{r \in \mathbb{R}^K} \left\{ \frac{1}{k+1} \sum_{\ell=1}^{} \left[^{k+1}(\Phi r)(x_i) - \alpha \max\{g(y_\ell), (\Phi r^{(k)})(y_\ell)\right]^2 \right\}$$

   (so we are using $(\Phi r^{(k)})(y_\ell)$ as an approximation for $Q^*(y)$).

As $k \to \infty$, $\|\Phi r - F\Phi r^{(k)}\|_\pi^2$. Again,

$$(FQ)(x) = \alpha \sum_{y \in S} P_{x,y} \max\{g(y), Q(y)\}.$$

With one sample we can get an unbiased estimate of $(FQ^*)(x)$. Does the iterate $Q := \Pi FQ$ converge? Yes.

Lemma 1.22.10.1 (Proposition 6.6.1 in Bertsekas [2012b], p. 506). F is a contraction mapping with respect to $\|\cdot\|_\pi$.

Proof. p. 506 of Bertsekas [2012b]. □

Question: given k, $(\Pi F)^k Q \to \tilde{Q}$, how close is $\tilde{Q}$ to Q^*? Note $\tilde{Q} \neq \Pi Q^*$. Rather, $\tilde{Q}$ is the fixed point of ΠF; $\tilde{Q} = \Pi F \tilde{Q}$.

Theorem 1.22.10.2.

$$\|\tilde{Q} - Q^*\|_\pi \leq \frac{1}{\sqrt{1-\alpha^2}} \|\Pi Q^* - Q^*\|_\pi.$$

Proof. Observe that $\tilde{Q} \in \operatorname{span} \Phi$. Also, $\Pi\tilde{Q} \in \operatorname{span} \Phi$. Therefore $\tilde{Q} - \Phi Q^* \perp \Pi Q^* - Q^*$ since $\Pi Q^* - Q^*$ is in the nullspace of Φ. Now

$$\begin{aligned} \|\tilde{Q} - Q^*\|_\pi^2 &= \|\tilde{Q} - \Pi Q^*_\Pi Q^* - Q^*\|_\pi^2 \\ &= \|\tilde{Q} - \Pi Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \\ &= \Pi F \tilde{Q} - \Pi F Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \\ \leq \alpha \|\tilde{Q} - Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \end{aligned}$$

because $\Pi Q^* = \Pi F Q^*$ and ΠF is a contraction mapping. Observe that

$$\|\tilde{Q} - Q^*\|_\pi \leq \frac{\|\Pi Q^* - Q^*\|_\pi^2}{\sqrt{1-\alpha^2}}.$$

□

proved error bound: $\|\hat{Q} - Q^*\|_\pi \leq \frac{1}{\sqrt{1-\alpha^2}} \cdot \|\Pi Q^* - Q^*\|_\pi$. This error bound highlights the importance of choosing a good approximation architecture. Φ has a span of K basis vectors; if it is close to Q^*, so is the projection of Q^* onto the span of Φ.

Question: how well does a policy generated from $\hat{Q}$ perform? An approach: let $\hat{J} := \max\{g, \hat{Q}\}$ and $\hat{J}$ will be our approximation of J^* (the optimal value function). Let μ be a greedy policy with respect to $\hat{J}$; that is, $T_\mu \hat{J} = T\hat{J}$. From our lecture in Week 6,

$$\begin{aligned}\|J^* - J_\mu\|_\infty &\leq \frac{2\alpha}{1-\alpha} \\ &\leq \frac{2\alpha}{1-\alpha}\|Q^* - \hat{Q}\|_\infty,\end{aligned}$$

but this distance can diverge to ∞ under $\|\cdot\|_\infty$. So we need a slightly different approach.

A new approach: $\hat{Q}$ is our approximation of Q^*, which is the optimal value of continuing on. Let $\hat{\tau}$ be a random variable for the stopping time derived from $\hat{Q}$; that is,

$$\hat{\tau} = \min\{t : \hat{Q}(x_t) \leq g(x_t)\}.$$

x_t is well-defined because we have one transition matrix (one Markov chain). This defines a policy (as long as approximate value of continuing is greater than stopping, keep going; once reversed, stop.) That is, the stopping region derived from $\hat{Q}$ is given by

$$\{x : \hat{Q}(x) \leq g(x)\}.$$

Recall again that x may in general be extremely high-dimensional. Note: the optimal stopping region is

$$\{x : Q^*(x) \leq g(x)\}.$$

The payoff associated with the policy derived from $\hat{Q}$ is given by

$$J_{\hat{\tau}}(x) = \mathbb{E}\left[\alpha^{\hat{\tau}} g(x_{\hat{\tau}}) \mid X_0 = x\right]$$

Let

$$\tau^* = \min\{t : Q^*(x_t) \leq g(x_t)\}.$$

denote the (actual) optimal stopping time. The optimal expected reward is

$$J^*(x) = \mathbb{E}\left[\alpha^{\tau^*} g(x_{\tau^*}) \mid X_0 = x\right].$$

Our goal is to compare J^* with $J_{\hat{\tau}}$.

Theorem 1.22.10.3.

$$\|J^* - J_{\hat{\tau}}\|_\pi \leq \frac{1}{1-\alpha}\|\hat{Q} - Q^*\|_\pi.$$

Remark 11. Observe that combining this result with the result from last time yields

$$\begin{aligned}\|J^* - J_{\hat{\tau}}\|_\pi &\leq \frac{1}{1-\alpha}\|\hat{Q} - Q^*\|_\pi \\ &\leq \frac{1}{(1-\alpha)\sqrt{1-\alpha^2}}\|\Pi Q^* - Q^*\|_\pi.\end{aligned}$$

Proof. Let $Q_{\hat{\tau}}(x)$ be the expected reward of the following policy: continue the process in the next step and follow the policy based on $\hat{\tau}$ thereafter.

We claim that $\|J^* - J_{\hat{\tau}}\|_\pi \leq \|Q^* - \hat{Q}\|_\pi + \|Q^* - Q_{\hat{\tau}}\|_\pi$. (Note that $Q_{\hat{\tau}} \neq \hat{Q}$. $\hat{Q}$ is derived from approximate value iteration; it is the unique fixed point of $\Pi F(\cdot)$. $Q_{\hat{\tau}}$ is the value of a policy.)

To prove this claim, note that

$$\begin{aligned}J^*(x) - J_{\hat{\tau}}(x) &= \begin{cases}\max\{g(x), Q^*(x)\} - g(x), & g(x) \geq \hat{Q}(x), \\ \max\{g(x), Q^*(x)\} - Q_{\hat{\tau}}(x), & g(x) < \hat{Q}(x).\end{cases} \\ &= \begin{cases}g(x) - g(x) = 0, & g(x) \geq \hat{Q}(x), g(x) \geq Q^*(x) \\ Q^*(x) - g(x), & g(x) \geq \hat{Q}(x), g(x) < Q^*(x) \\ Q^*(x) - Q_{\hat{\tau}}(x), & g(x) < \hat{Q}(x), g(x) < Q^*(x), \\ g(x) - Q_{\hat{\tau}}(x), & g(x) < \hat{Q}(x), g(x) \geq Q^*(x).\end{cases}\end{aligned}$$

Consider the cases one at at time.

1. $J^*(x) - J_{\hat{\tau}}(x) = g(x) - g(x) = 0$.
2. Since J^* is optimal, $0 \leq J^*(x) - J_{\hat{\tau}}(x) = Q^*(x) - g(x) \leq Q^*(x) - \hat{Q}(x)$ since by assumption of this case $\hat{Q}(x) \leq g(x)$.
3. We already have what we need: $0 \leq J^*(x) - J_{\hat{\tau}}(x) = Q^*(x) - Q_{\hat{\tau}}(x)$.
4. We have $Q^*(x) \leq g(x) < \hat{Q}(x)$ by assumption. Then

$$\begin{aligned}0 &\leq J^*(x) - J_{\hat{\tau}}(x) \\ &= g(x) - Q_{\hat{\tau}}(x) \\ &\leq \hat{Q}(x) - Q_{\hat{\tau}}(x) \\ &= \underbrace{\hat{Q}(x) - Q^*(x)}_{\geq 0 \text{ by assumption}} + \underbrace{Q^*(x) - Q_{\hat{\tau}}(x)}_{\geq 0 \text{ because } Q^* \text{ is optimal}}\end{aligned}$$

In all four cases, $\|J^* - J_{\hat{\tau}}\|_\pi \leq \|Q^* - \hat{Q}\|_\pi + \|Q^* - Q_{\hat{\tau}}\|_\pi$, proving the claim.

Now we want to use the claim to prove the theorem. We have

$$\begin{aligned}\|J^* - J_{\hat{\tau}}\|_\pi &\leq \|Q^* - \hat{Q}\|_\pi + \|Q^* - Q_{\hat{\tau}}\|_\pi \\ &= \|Q^* - \hat{Q}\|_\pi + \|\alpha P J^* - \alpha P J_{\hat{\tau}}\|_\pi \\ &\leq \|Q^* - \hat{Q}\|_\pi + \alpha \|J^* - J_{\hat{\tau}}\|_\pi\end{aligned}$$

where the second inequality followed because P is a nonexpansive mapping. Therefore

$$\begin{aligned}\|J^* - J_{\hat{\tau}}\|_\pi &\leq \frac{1}{1-\alpha}\|Q^* - \hat{Q}\|_\pi \\ &\leq \frac{1}{1-\alpha} \cdot \frac{1}{\sqrt{1-\alpha^2}}\|\Pi Q^* - Q^*\|_\pi.\end{aligned}$$

□

α may be very close to 1, in which case this proportion term can get quite large. We can do better than this. We can prove that

$$\|J^* - J_{\hat{\tau}}\|_\pi \leq \frac{2.17}{1-\alpha}\|\Pi Q^* - Q^*\|_\pi.$$

Lemma 1.22.10.4.

$$\|F\hat{Q} - \hat{Q}\|_\pi \leq (1+\alpha)\|Q^* - \Pi Q^*\|_\pi.$$

Proof. We claim that

$$\underbrace{(F\hat{Q} - \hat{Q})}_{\in \text{nullspace}(\Phi)} + \underbrace{(\Pi Q^* - Q^*)}_{\in \text{nullspace}(\Phi)} \perp \underbrace{\hat{Q} - \Pi Q^*}_{\in \text{span}(\Phi)},$$

where orthogonality is defined in terms of the π-weighted inner product. Recall that $\hat{Q}$ is the unique fixed point of ΠF. So $\hat{Q} = \Pi F \hat{Q}$, therefore $\hat{Q} \in \text{span}(\Phi)$. So ΠQ^* is in the span of Φ, so $Q^* - \Pi Q^*$ is in the nullspace of Φ. Because

$$\Pi(F\hat{Q} - \hat{Q}) = \Pi F\hat{Q} - \Pi\hat{Q} = \hat{Q} - \hat{Q} = 0.$$

Therefore

$$\|F\hat{Q} - Q^*\|_\pi^2 = \|F\hat{Q} - \hat{Q} + \Pi Q^* - Q^* + \hat{Q} - \Pi Q^*\|_\pi^2 \qquad = \|F\hat{Q} - \hat{Q} + \Pi Q^* - Q^*\|_\pi^2 + \|\hat{Q} - \Pi Q^*\|_\pi^2$$

where we used orthogonality of $F\hat{Q} - \hat{Q} + \Pi Q^* - Q^*$ and $\hat{Q} - \Pi Q^*$. Therefore

$$\begin{aligned}\|F\hat{Q} - \hat{Q} + \Pi Q^* - Q^*\|_\pi^2 = \|F\hat{Q} - \underbrace{Q^*}_{=FQ^*}\|_\pi^2 - \|\hat{Q} - \Pi Q^*\|_\pi^2 \qquad &\leq \alpha^2\|\hat{Q} - Q^*\|_\pi^2 - \alpha^2\|\hat{Q} - \Pi Q^*\|_\pi^2 \\ \leq \alpha^2\|\Pi Q^* - Q^*\|_\pi^2,\end{aligned}$$

where we used the formula $\|x\| - \|y\| \leq \|x - y\| \iff \|x\| \leq \|x - y\| + \|y\|$. Therefore

$$\|F\hat{Q} - Q^*\|_\pi \leq (1 + \alpha)\|\Pi Q^* - Q^*\|_\pi.$$

□

Lemma 1.22.10.5.

$$\|F\hat{Q} - Q_{\hat{\tau}}\|_\pi \leq \alpha\|\hat{Q} - Q_{\hat{\tau}}\|_\pi.$$

Proof. Let $\hat{V} := \max\{g, \hat{Q}\}$, and

$$V_{\hat{\tau}}(x) := \begin{cases} g(x), & g(x) \geq \hat{Q}(x), \\ Q_{\hat{\tau}}(x), & g(x) < \hat{Q}(x). \end{cases}$$

By definition,

$$F\hat{Q} = \alpha P \max\{g, \hat{Q}\} = \alpha P\hat{V},$$

$$Q_{\hat{\tau}} = \alpha P V_{\hat{\tau}},$$

so

$$\begin{aligned}\|F\hat{Q} - Q_{\hat{\tau}}\|_\pi &= \|\alpha P\hat{V} - \alpha P V_{\hat{\tau}}\|_\pi \\ &\leq \alpha\|\hat{V} - V_{\hat{\tau}}\|_\pi \\ &\leq \alpha\|\hat{Q} - Q_{\hat{\tau}}\|_\pi.\end{aligned}$$

□

Proof of new and improved error bound:

$$\begin{aligned}\|\hat{Q} - Q_{\hat{\tau}}\|_\pi &\leq \|\hat{Q} - F\hat{Q}\|_\pi + \|F\hat{Q} - Q_{\hat{\tau}}\|_\pi \\ &\leq \|\hat{Q} - F\hat{Q}\|_\pi + \alpha\|\hat{Q} - Q_{\hat{\tau}}\|_\pi\end{aligned}$$

where the second part followed by Lemma 2. **notes continue online**

SARSA: Q_0 in the span of our basis functions Φ.

1.23 Notes on Mathieu and Minsker [2019]

1.23.1 Notation

Let $(S, \mathcal{S})$ be a measurable space, and $X \in S$ is a random variable with distribution P. Let $\ell : \mathbb{R} \to \mathbb{R}_+$ be a loss function. We hope to find the function f in a class $\mathcal{F}$ of measurable functions from S to $\mathbb{R}$ minimizing the expected loss $\mathbb{E}\ell(f(X)) = \mathcal{L}(f) = P\ell(f)$. We assume the minimum is attained for a unique $f_* \in \mathcal{F}$.

The true distribution P is usually unknown, so we find a proxy for f_* using the empirical risk minimizer.

Definition 1.23.1 (**Empirical Risk Minimizer and Excess Risk**)**.** Let P_N be the empirical distribution of X based on the sample $X_1, \dots, X_N$. Define the empirical risk for a function f to be

$$\mathcal{L}_N(f) := P_N\ell(f) = \frac{1}{N}\sum_{j=1}^{N} \ell(f(X_j)).$$

Then the empirical risk minimizer is

$$\tilde{f}_N := \arg\min_{f \in \mathcal{F}} \mathcal{L}_N(f). \tag{1.60}$$

Performance of $f \in \mathcal{F}$ is measured via the **excess risk** $\mathcal{E}(f) := P\ell(f) - P\ell(f_*) = \mathbb{E}\left[\ell(f(X)) - \ell(f_*(X))\right]$. The excess risk of $\tilde{f}_N$ is a random variable

$$\mathcal{E}(f) := P\ell(\tilde{f}_N) - P\ell(f_*) = \mathbb{E}\left[\ell\left(\tilde{f}(X)\right) \mid X_1, \dots, X_N\right] - \mathbb{E}\ell(f_*(X)).$$

Example 1.23.1 (Regression)**.**

- $X = (Z, Y) \in \mathbb{R}^d \times \mathbb{R}$.
- $f(Z, Y) = Y - g(Z)$ for some g in a class $\mathcal{G}$ (such as the class of linear functions from $\mathbb{R}^d \to \mathbb{R}$).
- $\ell(x) = x^2$.
- $f_*(z, y) = y - g_*(z)$, where $g_*(z) = \mathbb{E}(Y \mid Z = z)$.
- $\mathcal{L}_N(f) = \frac{1}{N}\sum_{j=1}^{N} \ell(f(X_j)) = \frac{1}{N}\sum_{j=1}^{N}[Y_j - g(Z_j)]^2$.
- $\tilde{f}_N = \arg\min_{f \in \mathcal{F}} \left\{\frac{1}{N}\sum_{j=1}^{N}(f(X_j))^2\right\} = Y - \tilde{g}_N(Z) \quad \text{where} \quad \tilde{g}_N := \arg\min_{g \in \mathcal{G}} \left\{\frac{1}{N}\sum_{j=1}^{N}(Y_j - g(Z_j))^2\right\}$.
-

$$\begin{aligned}
\mathcal{E}(f) &= P\ell(\tilde{f}_N) - P\ell(f_*) \\
&= \mathbb{E}\left[\ell\left(\tilde{f}(X)\right) \mid X_1, \dots, X_N\right] - \mathbb{E}\ell(f_*(X)) \\
&= \mathbb{E}\left[(Y - \tilde{g}_N(Z))^2 \mid X_1, \dots, X_N\right] - \mathbb{E}\left[Y - \mathbb{E}(Y \mid Z)\right]^2
\end{aligned}$$

Definition 1.23.2. For two sequences $\{a_j\}_{j\geq 1} \subset \mathbb{R}$ and $\{b_j\}_{j\geq 1} \subset \mathbb{R}$ for $j \in \mathbb{N}$, the expression $a_j \lesssim b_j$ means there exists a constant $c > 0$ such that $a_j \leq cb_j$ for all $j \in \mathbb{N}$. $a_j \asymp b_j$ means that $a_j \lesssim b_j$ and $b_j \lesssim a_j$.

For a function $h : \mathbb{R}^d \to \mathbb{R}$, the authors define

$$\arg\min_{y\in\mathbb{R}^d} h(y) := \left\{y \in \mathbb{R}^d : h(y) \leq h(x) \quad \forall x \in \mathbb{R}^d\right\}.$$

Also, $\|h\|_\infty := \operatorname{ess\,sup}\left\{|h(y)| : y \in \mathbb{R}^d\right\}$. $L(h)$ will stand for a Lipschitz constant of h. For $f \in \mathcal{F}$, let $\sigma^2(\ell, f) := \operatorname{Var}(\ell(f(X))$, and for any subset $\mathcal{F}' \subseteq \mathcal{F}$, denote $\sigma^2(\ell, \mathcal{F}') := \sup_{f\in\mathcal{F}'} \sigma^2(\ell, f)$.

1.23.2 Section 1

The focus of this paper is on the situation when marginal distributions of the process $\{\ell(f(X)), f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ are allowed to be heavy-tailed in the sense that only their first 2 to 4 moments are finite. The authors also consider a framework of *adversarial contamination*, when the initial data set of cardinality N is merged with a set of $\mathcal{O} < N$ outliers generated by an adversary who has an opportunity to inspect the data. The combined data set of cardinality $N^\circ = N + \mathcal{O}$ is presented to the algorithm. (The authors assume an upper bound for proportion of contamination $\mathcal{O}/N$ is known.)

The robust estimator the authors will propose incorporates ideas from a "median-of-means" estimator as well as Catoni's estimator [Catoni, 2012], which relies on truncation of the data.

Definition 1.23.3 (Median of means estimator; from Devroye et al. [2016], Section 4.1). Let b be a positive integer and let $x_1^b \in \mathbb{R}^b$. Let $q_{1/2}$ denote the median of the numbers $x_1, \ldots, x_b$; that is,

$$q_{1/2}\left(x_1^b\right) = x_i \quad \text{where } |\{k \in [b] : x_k \leq x_i\}| \geq \frac{b}{2} \text{ and } |\{k \in [b] : x_k \geq x_i\}| \geq \frac{b}{2}.$$

(If more than one i fit the above description, take the smallest one.) Let $\delta \in [e^{1-/n2}, 1)$. Choose $b = \lceil \log(1/\delta) \rceil$ (note that $b \leq n/2$). Now divide $[n]$ into b blocks (disjoint subsets) B_i, $1 \leq i \leq b$, each of size $|B_i| \geq k = \lfloor n/b \rfloor \geq 2$. Given $x_1^n \in \mathbb{R}^n$, define

$$y_{n,\delta,i}\left(x_1^n\right) := \frac{1}{|B_i|} \sum_{j\in B_i} x_i,$$

$$y_{n,\delta}\left(x_1^n\right) := \left(y_{n,\delta,i}\left(x_1^n\right)\right)_{i=1}^b \in \mathbb{R}^b,$$

and define the median-of-means estimator by

$$\hat{E}_{n,\delta}\left(x_1^n\right) := q_{1/2}\left(y_{n,\delta}\left(x_1^n\right)\right).$$

Definition 1.23.4 (Robust mean estimators). Let $k \leq N$ be an integer. Assume $G_1, \ldots, G_k$ are disjoint subsets of the index set $[N]$ of cardinality $|G_j| = n \geq \lfloor N/k \rfloor$ each. Given $f \in F$, let

$$\overline{\mathcal{L}}_j(f) := \frac{1}{n} \sum_{i \in G_j} \ell(f(X_i))$$

be the empirical mean evaluated over the subsample indexed by G_j. Given a convex, even function $\rho : \mathbb{R} \to \mathbb{R}_+$ and $\Delta > 0$, set

$$\begin{aligned}\hat{\mathcal{L}}^{(k)}(f) &:= \arg\min_{y \in \mathbb{R}} \left\{ \sum_{j=1}^{k} \rho\left(\sqrt{n} \cdot \frac{\overline{\mathcal{L}}_j(f) - y}{\Delta} \right) \right\} \\ &= \arg\min_{y \in \mathbb{R}} \left\{ \sum_{j=1}^{k} \rho\left(\sqrt{n} \cdot \frac{n^{-1} \sum_{i \in G_j} \ell(f(X_i)) - y}{\Delta} \right) \right\}\end{aligned}$$

Remark 12. Note that if $\rho(x) = x^2$, $\hat{\mathcal{L}}^{(k)}(f)$ is equal to the sample mean:

$$\begin{aligned}\hat{\mathcal{L}}^{(k)}(f) &= \arg\min_{y \in \mathbb{R}} \left\{ \sum_{j=1}^{k} \left(\sqrt{n} \cdot \frac{n^{-1} \sum_{i \in G_j} \ell(f(X_i)) - y}{\Delta} \right)^2 \right\} \\ &= \arg\min_{y \in \mathbb{R}} \left\{ \frac{n}{\Delta^2} \sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i \in G_j} \ell(f(X_i)) - y \right)^2 \right\} \\ &= \arg\min_{y \in \mathbb{R}} \left\{ \sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i \in G_j} \ell(f(X_i)) - y \right)^2 \right\} \\ &= \arg\min_{y \in \mathbb{R}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\ell(f(X_i)) - y \right)^2 \right\} \\ &= \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i)).\end{aligned}$$

If $\rho(x) = |x|$, it turns out that $\hat{\mathcal{L}}^{(k)}(f)$ is the median-of-means estimator:

$$\begin{aligned}\hat{\mathcal{L}}^{(k)}(f) &= \arg\min_{y \in \mathbb{R}} \left\{ \sum_{j=1}^{k} \left| \sqrt{n} \cdot \frac{n^{-1} \sum_{i \in G_j} \ell(f(X_i)) - y}{\Delta} \right| \right\} \\ &= \hat{E}_{N,\delta}\left(\ell(f(X)) \right).\end{aligned}$$

This paper focuses on the situation when ρ is similar to Huber's loss (ρ' is bounded and Lipschitz continuous). It is instructive to consider two cases. First, when $k = N$ (so that $n = 1$) and

$$\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))} \sqrt{N},$$

$\hat{\mathcal{L}}^{(k)}(f)$ is akin to Catoni's estimator. When n is large and

$$\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))},$$

we recover the "median-of-mean"-type estimator. (The standard median-of-means estimator corresponds to $\rho(x) = x$ and can be seen as a limit of $\hat{\mathcal{L}}^{(k)}(f)$ when $\Delta \to 0$; this case is not covered by results of the paper, as the authors require that ρ' is smooth and Δ is bounded from below.)

We can also construct an estimator that does not depend on the specific choice of subgroups $G_1, \ldots, G_k$.

Definition 1.23.5 (**Permutation-invariant robust mean estimator**)**.** Define

$$\mathcal{A}_N^{(n)} := \{J : J \subseteq [N], |J| = n\}.$$

Let h be a measurable, permutation-invariant function of n variables. Recall that a U-statistic of order n with kernel h based on an i.i.d. sample $X_1, \ldots, X_N$ is defined as

$$U_{N,n} := \frac{1}{\binom{N}{n}} \sum_{J \in \mathcal{A}_N^{(n)}} h\left(\{X_j\}_{j \in J}\right)$$

(see Section **??**). Given $J \in \mathcal{A}_N^{(n)}$, let

$$\overline{\mathcal{L}}(f; J) := \frac{1}{n} \sum_{i \in J} f(X_i).$$

(Note that $\overline{\mathcal{L}}(f; J)$ is a permutation-invariant function of its arguments.) Consider U-statistics of the form

$$U_{N,n}(z; f) = \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(\sqrt{n} \cdot \frac{\overline{\mathcal{L}}(f; J) - z}{\Delta}\right) = \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(\sqrt{n} \cdot \frac{n^{-1} \sum_{i \in J} f(X_i) - z}{\Delta}\right).$$

Then the permutation-invariant version of $\hat{\mathcal{L}}^{(k)}(f)$ is naturally defined as

$$\hat{\mathcal{L}}_U^{(k)}(f) := \arg\min_{z \in \mathbb{R}} U_{N,n}(z; f).$$

Assuming that $\hat{\mathcal{L}}^{(k)}(f)$ provides good approximation of the expected loss $\mathcal{L}(f)$ of each individual $f \in \mathcal{F}$, it is natural to consider

$$\hat{f}_N := \arg\min_{f \in \mathcal{F}} \hat{\mathcal{L}}^{(k)}(f)$$

as well as its permutation-invariant analogue

$$\hat{f}_N^U := \arg\min_{f \in \mathcal{F}} \hat{\mathcal{L}}_U^{(k)}(f)$$

as alternatives to standard risk minimization (1.60). The goal of this paper is to get general bounds for the excess risk of the estimators $\hat{f}_N$ and $\hat{f}_N^U$ under minimal assumptions on the stochastic process $\{\ell(f(X)), f \in \mathcal{F}\}$.

1.24 Random Forests and Notes on Chi et al. [2020]

Definition 1.24.1. A real- or complex-valued function f on a d-dimensional Euclidean space is **Holder continuous** if there exist nonnegative real constants C and α such that

$$|f(x) - f(y)| \leq C\|x - y\|^\alpha$$

for all x and y in the domain of f. (This condition can be formulated for functions between any two metric spaces.) The number α is called the **exponent** of the Holder condition. A function on an interval satisfying the condition with $\alpha > 1$ is constant. Note that if $\alpha = 1$ then the function is Lipschitz continuous, and for $\alpha \in (0, 1)$ Holder continuity is weaker than Lipschitz continuity. For any $\alpha > 0$ the condition implies the function is uniformly continuous.

1.24.1 Section 2 (Terminology and Review of Random Forest)

Definition 1.24.2. Let $\mathcal{T}$ be the set of rectangles $\boldsymbol{t} = \times_{j=1}^p t_j \subset \boldsymbol{t}_0 := [0, 1]^p$, where each t_j is a closed or half-closed interval in $[0, 1]$. A **cell** or **subcell** is an element of $\mathcal{T}$. A **cut** or **split** is a feature and location pair that can be used for separating the parent cell. That is, for a nonempty cell $\boldsymbol{t}$, a cut is a pair (s, c) with $s \in \{1, \ldots, p\}$ and $c \in t_s$, and the **daughter cells** obtained by separating $\boldsymbol{t}$ accordingly are

$$\times_{j=1}^{s-1} t_j \times (t_s \cap [0, c)) \times_{j=s+1}^p t_j$$

and

$$\times_{j=1}^{s-1} t_j \times (t_s \cap [c, 1]) \times_{j=s+1}^p t_j.$$

Observe that the daughter cells of an empty cell are two empty cells. In practice, there will be a restriction of available directions (features) on a cut. At an initial level (i.e., level 0), we have a root cell $\boldsymbol{t}_0 = [0, 1]^p$. Then two subcells are obtained at level 1 after a cut for the root cell. Then we cut those two subcells to create 4 subcells, and so on. Thus at each level ℓ, when growing a tree there are 2^ℓ subcells to be cut.

Definition 1.24.3 (A set of level k daughter cell paths)**.** Let $\mathcal{D} \subset \mathcal{T}^k$ be a set composed of 2^k k-tuples (that is, $\#\mathcal{D} = 2^k$). $\mathcal{D}$ is a **set of level k daughter cell paths** (also called a **tree**) if and only if

1. For each $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in \mathcal{D}$, $\boldsymbol{t}_\ell \in \mathcal{T}$ is a subset of $[0, 1]^p$ and in particular is one of the daughter cells of $\boldsymbol{t}_{\ell-1}$ for each $\ell \in [k]$, and
2. The set of unique elements in $\{t_\ell : (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in \mathcal{D}\}$ is a partition of $\boldsymbol{t}_0$ for each $\ell \in [k]$.

(That is, for any $\ell \in [k]$, the set of unique elements in the ℓ^{th} slot of an element of $\mathcal{D}$ form a partition of $[0,1]^p$. So for example, the set of unique first entries of elements of this set are two non-overlapping hyperrectangles whose union is $[0,1]^p$, and the unique last entries of elements of this set form a partition of $[0,1]^p$ into 2^k non-overlapping hyperrectangles.)

Denote by $\tilde{\mathcal{D}}$ the set of all such sets $\mathcal{D}$.

Note that a set of level k daughter cell paths $\mathcal{D} \in \tilde{\mathcal{D}}$ has exactly 2^k tuples. Also, as long as there are no empty daughter cells at each level, these 2^k tuples are mutually exclusive.

The random forest algorithm usually puts a restriction on the available feature subset when deciding on the feature for each cut. Denote by $\Theta_k \in \mathcal{P}(\{1,\dots,p\}) = \mathcal{P}([p])$ (where $\mathcal{P}([p])$ denotes the power set of $[p]$) the constraint for each of the 2^{k-1} cuts at level k. Given a set of observations, a sequence of constraints $\{\Theta_i\}_{i=1}^{\infty}$, and some positive integer k, the random forest algorithm produces a set of level k daughter cell paths (a tree). Most random forest implementations use the CART-split criterion for growing the tree. Then another tree is created using the same set of observations and the same positive integer k but a different sequence of constraints.

Definition 1.24.4 (Tree growing rule)**.** A **tree growing rule** denoted as $T : \mathbb{N} \times (\mathcal{P}([p]))^k \to \tilde{\mathcal{D}}$ is a mapping that takes some positive integer k and k feature subset constraints as inputs and outputs a set of level k daughter cell paths (a tree).

Denote by $\#S$ the number of elements in a set S. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a given sample with $\boldsymbol{x}_i := (x_{i1},\dots,x_{ip})^\top \in [0,1]^p$ the p-dimensional random covariate vector and $y_i \in \mathbb{R}$ the response. We define the summation over an empty set as zero.

Definition 1.24.5 (Sample CART-split criterion)**.** Given a cell $\boldsymbol{t} \in \mathcal{T}$ and a feature subset $\Theta \in \mathcal{P}([p])$, the sample CART-split criterion is defined as

$$(\hat{j}, \hat{c}) := \underset{j \in \Theta, c \in \{x_{ij} : \boldsymbol{x}_i \in \boldsymbol{t}\}}{\arg\min} \left\{ \sum_{i \in \{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} < c\}} (\overline{y}_\ell(\boldsymbol{t}, c) - y_i)^2 + \sum_{i \in \{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} \geq c\}} (\overline{y}_r(\boldsymbol{t}, c) - y_i)^2 \right\}$$

where

$$\overline{y}_\ell(\boldsymbol{t}, c) := \sum_{i \in \{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} < c\}} \frac{y_i}{\#\{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} < c\}}, \qquad \overline{y}_r(\boldsymbol{t}, c) := \sum_{i \in \{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} \geq c\}} \frac{y_i}{\#\{i : \boldsymbol{x}_i \in \boldsymbol{t}, x_{ij} \geq c\}}$$

The above optimization breaks ties evenly. For a subcell $\boldsymbol{t} \subset [0,1]^p$ with $\#\{i : \boldsymbol{x}_i \in \boldsymbol{t}\} = 0$, a random cut is optimal.

Definition 1.24.6 (Sample tree growing rule; Definition 3 in Chi et al. [2020])**.** For each positive integer k, a subset of sample indices $\mathcal{A} \in \mathcal{P}([n])$, and feature subsets $\Theta_1,\dots,\Theta_k$, the sample tree growing rule

$$\hat{T}_{(n,\mathcal{A})} : (\mathcal{P}([p]))^k \to \tilde{\mathcal{D}}$$

is such that if $(\boldsymbol{t}_1,\dots,\boldsymbol{t}_k) \in \hat{T}_{(n,\mathcal{A})}(\Theta_1,\dots,\Theta_k)$, then for each $\ell \in [k]$, $\boldsymbol{t}_\ell$ is one of the daughter cells of $\boldsymbol{t}_{\ell-1}$ constructed by the sample CART-split criterion with the sample given in $\mathcal{A}$.

Next we define a level k sample forest model; informally, it is the average of many tree models. Denote by $\mathcal{X}_n := (\boldsymbol{x}_1^\top, y_1, \ldots, \boldsymbol{x}_n^\top, y_n)^\top \in \{[0,1]^p \times \mathbb{R}\}^n$.

Definition 1.24.7 (Level k sample forest model)**.** Let $\tilde{\mathcal{A}} \in \mathcal{P}([n]) \times \ldots \times \mathcal{P}([n])$ be a set of sample indices such that the elements in $\tilde{\mathcal{A}}$ do not repeat and are of the same predetermined size; that is, for each $\mathcal{A} \in \tilde{\mathcal{A}}$, it holds that $|\mathcal{A}| = \lceil bn \rceil$ for some $b \in (0,1]$. Denote by $\hat{m}_{k,T,\mathcal{A}} : (\mathcal{P}([p]))^k \times [0,1]^p \times \{[0,1]^p \times \mathbb{R}\}^n \to \mathbb{R}$ the tree model such that for each $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in T(\Theta_1, \ldots, \Theta_k)$ and each $\boldsymbol{c} \in \boldsymbol{t}_k$,

$$\hat{m}_{k,T,\mathcal{A}}(\Theta_1, \ldots, \Theta_k, \boldsymbol{c}, \mathcal{X}_n) = \sum_{i \in (\{i : \boldsymbol{x}_i \in \boldsymbol{t}_k\} \cap \mathcal{A})} \frac{y_i}{\#(\{i : \boldsymbol{x}_i \in \boldsymbol{t}_k\} \cap \mathcal{A})}.$$

Thus the forest model considering random sampling given $\boldsymbol{c} \in [0,1]^p$, sample $\mathcal{X}_n$, and feature constraints is defined as

$$\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k,T,\mathcal{A}}(\Theta_1, \ldots, \Theta_k, \boldsymbol{c}, \mathcal{X}_n).$$

Since standard random forest packages draw random feature subsets as constraints, denote by $\{\boldsymbol{\Theta}_i\}_{i=1}^\infty$ a sequence of random constraints (that is, in the notation we have previously used, $\{\Theta_i\}_{i=1}^\infty$ is a realization of $\{\boldsymbol{\Theta}_i\}_{i=1}^\infty$).

Definition 1.24.8. The forest prediction (using both random feature subsets and random sampling) given T, $\boldsymbol{X}$ (a random variable taking on values in $[0,1]^p$), and $\mathcal{X}_n$ is defined as

$$\mathbb{E}\left(\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k,T,\mathcal{A}}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n) \mid \boldsymbol{X}, \mathcal{X}_n \right).$$

For the special forest model with $\#\tilde{\mathcal{A}} = 1$ and $\mathcal{A} \in \tilde{\mathcal{A}}$ the full sample indices, we use the following notation:

$$\hat{m}_{k,T}(\Theta_1, \ldots, \Theta_k, \boldsymbol{c}, \mathcal{X}_n) := \frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k,T,\mathcal{A}}(\Theta_1, \ldots, \Theta_k, \boldsymbol{c}, \mathcal{X}_n). \tag{1.61}$$

We call this model a **tree model**. It is worth mentioning that both sample forest and tree models are averaged over all possible feature subset constraints. Then the (level k) random forest prediction of a standard random forest algorithm at the test point $\boldsymbol{X}$ given $\mathcal{X}_n$ is defined as

$$\mathbb{E}\left(\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k,\hat{T}_{(n,\mathcal{A})},\mathcal{A}}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n) \mid \boldsymbol{X}, \mathcal{X}_n \right), \tag{1.62}$$

where $\tilde{\mathcal{A}}$ is an arbitrary given set.

Remark 13. There are two layers of randomness in the random forest model: the random feature subsets and random sampling. For the random forest prediction in (1.62), both kinds of randomness are taken into account; in particular, all possible feature subsets are included. For practical implementation, it is not required to consider all possible feature subsets. The standard random forest R package grows $B \leq \#A$ trees each of which involves random resampling and feature subsets of size $\lceil \gamma_0 p \rceil$.

1.24.2 Section 3: Approximation Accuracy

Definition 1.24.9. Denote $\mathbb{P}(\cdot \mid \boldsymbol{X} \in \boldsymbol{t})$ as $\mathbb{P}_{\boldsymbol{t}}(\cdot)$, and similarly denote $\mathbb{E}(\cdot \mid \boldsymbol{X} \in \boldsymbol{t})$ as $\mathbb{E}_{\boldsymbol{t}}(\cdot)$. Given a nonempty cell $\boldsymbol{t} \in \mathcal{T}$ and a feature subset $\Theta \subset \{1, \ldots, p\}$, the **best cut** according to the **population CART-split criterion** is defined as

$$(j^*, c^*) := \underset{j \in \Theta, c \in \boldsymbol{t}_j}{\arg\inf} \left\{ \mathbb{P}_{\boldsymbol{t}}(X_j < c) \mathrm{Var}\left(m \mid \{X_j < c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right) + \mathbb{P}_{\boldsymbol{t}}(X_j \geq c) \mathrm{Var}\left(m \mid \{X_j \geq c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right) \right\}.$$

Ties are broken randomly. If $\boldsymbol{t}$ is empty, both of its daughter cells are empty. We define the **best constrained (unconstrained) daughter cells** as the daughter cells resulting from the optimal cut when Θ is a nontrivial (trivial) constraint.

Definition 1.24.10. The **impurity decrease** of a cut (j, c) given $\boldsymbol{t}$ is defined to be

$$\mathrm{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) - \left[\mathbb{P}_{\boldsymbol{t}}(X_j < c) \mathrm{Var}\left(m \mid \{X_j < c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right) + \mathbb{P}_{\boldsymbol{t}}(X_j \geq c) \mathrm{Var}\left(m \mid \{X_j \geq c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right) \right]. \tag{1.63}$$

Note that the population CART-split criterion from Definition 1.24.9 maximizes the impurity decrease (and can be equivalently defined this way).

Definition 1.24.11 (Population tree growing rule; Definition 4 in Chi et al. [2020])**.** Let the regression function and joint distribution of covariates be given. For each positive integer k and feature subsets $\Theta_1, \ldots, \Theta_k$, the **population tree growing rule** T^* is such that if $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in T^*(\Theta_1, \ldots, \Theta_k)$, then for each $\ell \in [k]$, $\boldsymbol{t}_\ell$ is one of the daughter cells of $\boldsymbol{t}_{\ell-1}$ constructed by the population CART-split criterion.

Given a tree growing rule T, a level k population tree model is defined as a function $m^*_{k,T}$ of the feature subset constraints $\Theta_1, \ldots, \Theta_k$ and a p-dimensional vector such that for each $\boldsymbol{c} \in \boldsymbol{t}_k$ and $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in T(\Theta_1, \ldots, \Theta_k)$,

$$m^*_{k,T}(\Theta_1, \ldots, \Theta_k, \boldsymbol{c}) = \mathbb{E}(m \mid \boldsymbol{X} \in \boldsymbol{t}_k). \tag{1.64}$$

The prediction at a test point $\boldsymbol{X}$ is then given by

$$\mathbb{E}\left(m^*_{k,T}(\Theta_1, \ldots, \Theta_k, \boldsymbol{X}) \mid \boldsymbol{X}\right).$$

Also, the following variance decomposition formula is useful:

Lemma 1.24.2.1 (Variance decomposition formula; Lemma 1 in Chi et al. [2020])**.** Let $\boldsymbol{t} \in [0,1]^p$ be a given cell, and let $\boldsymbol{t}'$ and $\boldsymbol{t}''$ be two daughter cells of $\boldsymbol{t}$. The the conditional variance of $m := m(\boldsymbol{X})$ on $\boldsymbol{t}$ admits the following decomposition:

$$\begin{aligned} \mathrm{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) = \; & \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}') \mathrm{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}') + \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}'') \mathrm{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}'') \\ & + \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}') \left(\mathbb{E}_{\boldsymbol{t}'}(m) - \mathbb{E}_{\boldsymbol{t}}(m)\right)^2 + \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}'') \left(\mathbb{E}_{\boldsymbol{t}''}(m) - \mathbb{E}_{\boldsymbol{t}}(m)\right)^2. \end{aligned} \tag{1.65}$$

For notational convenience, define

$$(I)_{\boldsymbol{t},\boldsymbol{t}'} := \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}')\text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}') + \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}'')\text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}'')$$

and

$$(II)_{\boldsymbol{t},\boldsymbol{t}'} := \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}')\left(\mathbb{E}_{\boldsymbol{t}'}(m) - \mathbb{E}_{\boldsymbol{t}}(m)\right)^2 + \mathbb{P}_{\boldsymbol{t}}(\boldsymbol{X} \in \boldsymbol{t}'')\left(\mathbb{E}_{\boldsymbol{t}''}(m) - \mathbb{E}_{\boldsymbol{t}}(m)\right)^2$$

so we can write (1.65) as

$$\text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) = (I)_{\boldsymbol{t},\boldsymbol{t}'} + (II)_{\boldsymbol{t},\boldsymbol{t}'}.$$

Note that since $\boldsymbol{t} = \boldsymbol{t}' \cup \boldsymbol{t}''$, $(I)_{\boldsymbol{t},\boldsymbol{t}'} = (I)_{\boldsymbol{t},\boldsymbol{t}''}$ and $(II)_{\boldsymbol{t},\boldsymbol{t}'} = (II)_{\boldsymbol{t},\boldsymbol{t}''}$. Note also that the population CART-split criterion from Definition 1.24.9 minimizes $(I)_{\boldsymbol{t},\boldsymbol{t}'}$ for the case of axis-aligned cuts. Note also that in the case of axis-aligned cuts, the impurity decrease formula (1.63) can be written as

$$\begin{aligned}
&\text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) - [\mathbb{P}_{\boldsymbol{t}}(X_j < c)\text{Var}\left(m \mid \{X_j < c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right) + \mathbb{P}_{\boldsymbol{t}}(X_j \geq c)\text{Var}\left(m \mid \{X_j \geq c\} \cap \{\boldsymbol{X} \in \boldsymbol{t}\}\right)] \\
&= \text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) - (I)_{\boldsymbol{t},\boldsymbol{t}'} \\
&= (II)_{\boldsymbol{t},\boldsymbol{t}'}.
\end{aligned}$$

Therefore the population CART-split criterion from Definition 1.24.9 maximizes $(II)_{\boldsymbol{t},\boldsymbol{t}'}$. Also, zero impurity on $\boldsymbol{t}$ means that the regression function on $\boldsymbol{t}$ is just an intercept.

Technical Conditions

The below conditions are needed for this paper's theoretical results.

1. **Condition 1: Sufficient Impurity Decrease (SID).** There exist some $\alpha_1 \geq 1$ and $q_1 \geq 1$ such that for each cell $\boldsymbol{t}$, $\text{Var}\left(m \mid \boldsymbol{X} \in \boldsymbol{t}\right) < \alpha_1 \left((II)_{\boldsymbol{t},\boldsymbol{t}^*}\right)^{1/q_1}$, where $\boldsymbol{t}^*$ is one of the best unconstrained daughter cells of $\boldsymbol{t}$ (recall Definition 1.24.9).

   - For the specific case of $q_1 = 1$, Condition 1 requires that there is a minimum impurity decrease rate (i.e., the inverse of $\alpha_1 - 1$) for each cell when the cell is cut by the corresponding optimal cut. (Note that since $\text{Var}\left(m \mid \boldsymbol{X} \in \boldsymbol{t}\right) = (I)_{\boldsymbol{t},\boldsymbol{t}'} + (II)_{\boldsymbol{t},\boldsymbol{t}'}$, if $q_1 = 1$,

   $$\begin{aligned}
   \text{Var}\left(m \mid \boldsymbol{X} \in \boldsymbol{t}\right) < \alpha_1 (II)_{\boldsymbol{t},\boldsymbol{t}^*} \quad &\iff \quad \text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) - (I)_{\boldsymbol{t},\boldsymbol{t}'} = (II)_{\boldsymbol{t},\boldsymbol{t}'} < \alpha_1 (II)_{\boldsymbol{t},\boldsymbol{t}^*} - (I)_{\boldsymbol{t},\boldsymbol{t}'} \\
   &\iff \quad (I)_{\boldsymbol{t},\boldsymbol{t}'} < (\alpha_1 - 1)\,(II)_{\boldsymbol{t},\boldsymbol{t}^*} \\
   &\iff \quad (II)_{\boldsymbol{t},\boldsymbol{t}^*} > \frac{(I)_{\boldsymbol{t},\boldsymbol{t}'}}{\alpha_1 - 1}.
   \end{aligned}$$

   That is, since $(II)_{\boldsymbol{t},\boldsymbol{t}^*}$ is the impurity decrease, the impurity decrease has to exceed a certain threshold. If $q_1 > 1$ and $\alpha_1 > (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{\frac{q_1 - 1}{q_1}}$:

$$\begin{aligned}
\text{Var}\left(m \mid \boldsymbol{X} \in \boldsymbol{t}\right) < \alpha_1 (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{1/q_1} \quad &\iff \quad \text{Var}(m \mid \boldsymbol{X} \in \boldsymbol{t}) - (I)_{\boldsymbol{t},\boldsymbol{t}'} = (II)_{\boldsymbol{t},\boldsymbol{t}'} < \alpha_1 (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{1/q_1} - (I)_{\boldsymbol{t},\boldsymbol{t}'} \\
&\iff \quad (I)_{\boldsymbol{t},\boldsymbol{t}'} < \left(\alpha_1 - (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{\frac{q_1-1}{q_1}}\right) (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{1/q_1} \\
&\iff \quad (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{1/q_1} > \frac{1}{\alpha_1 - (II)_{\boldsymbol{t},\boldsymbol{t}^*}^{\frac{q_1-1}{q_1}}} (I)_{\boldsymbol{t},\boldsymbol{t}'}
\end{aligned}$$

2. **Condition 2.** Assume that $|m| \leq M_0$ for some $M_0 > 0$.

3. **Condition 3.** Consider a tree growing rule T such that T depends only on the feature subset constraint and there exists some positive integer k, $\epsilon > 0$, and $\alpha_2 \geq 1$ such that for each $\ell \in [k]$, feature subset constraint $\Theta_1, \ldots, \Theta_k$, and $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k) \in T(\Theta_1, \ldots, \Theta_k)$,

   (a) $(II)_{\boldsymbol{t}_{\ell-1},\boldsymbol{t}_\ell} \leq \epsilon$ only if $(II)_{\boldsymbol{t}_{\ell-1},\boldsymbol{t}_\ell^*} \leq \alpha_2 \epsilon$, and

   (b) If $(II)_{\boldsymbol{t}_{\ell-1},\boldsymbol{t}_\ell} > \epsilon$, then $(II)_{\boldsymbol{t}_{\ell-1},\boldsymbol{t}_\ell^*} \leq \alpha_2 (II)_{\boldsymbol{t}_{\ell-1},\boldsymbol{t}_\ell}$,

   where $\boldsymbol{t}_\ell^*$ is one of the best daughter cells of $\boldsymbol{t}_{\ell-1}$ given the feature subset constraint.

Main Results

Theorem 1.24.2.2 (Approximation accuracy; Theorem 1 in Chi et al. [2020])**.** Assume that Conditions 1 - 2 hold and the tree growing rule T satisfies item 1 of Condition 3 with some positive integer k, $\epsilon > 0$, and $\alpha_2 \geq 1$. Then

$$\mathbb{E}\left[m(\boldsymbol{X}) - m_{k,T}^*(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X})\right]^2 \leq \alpha_1 \left(\alpha_2 \epsilon\right)^{1/q_1} + \left(1 - \gamma_0 \left(\frac{\epsilon}{\epsilon + M_0^2}\right)\right)^k M_0^2,$$

where $m_{k,T}^*$ is the level k population tree model from (1.64) in Definition 1.24.11 and γ_0 is as described in Remark 13. Moreover, when item 2 of Condition 3 is satisfied and $q_1 = 1$, it holds that

$$\mathbb{E}\left[m(\boldsymbol{X}) - m_{k,T}^*(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X})\right]^2 \leq \alpha_1 \alpha_2 \epsilon + \left(1 - \frac{\gamma_0}{\alpha_1 \alpha_2}\right)^k M_0^2.$$

Theorem 1.24.2.2 provides a general approximation theory on the rate of convergence for random forest without assuming that the covariates are independent. This is the "noiseless case."

1.24.3 Consistency Rates (Section 4 of Chi et al. [2020])

Chi et al. introduce two additional regularity conditions.

- **Condition 4.** Assume that the distribution of covariates $\boldsymbol{X}$ has a density function f that is bounded away from 0 and ∞.
- **Condition 5.** Assume that $y_i = m(\boldsymbol{x}_i) + \epsilon_i$ with $\boldsymbol{x}_i$ i.i.d. realizations from $\boldsymbol{X}$ and ϵ_i i.i.d. with mean zero and a symmetric distribution, and $p = \mathcal{O}(n^K)$ for some positive constant K.

These are some basic assumptions in nonparametric regression models. In particular, this allows for polynomially growing dimension p. Some key lemmas follow.

Lemma 1.24.3.1 (Estimation error; Lemma 2 in Chi et al. [2020])**.** Assume that Conditions 2 and 5 hold, $0 < \eta < 1/2$, $0 < c < (\log 2)^{-1}(1/2 - \eta)$, and $\mathbb{E}|\epsilon_1|^q < \infty$ for sufficiently large q. Then it holds that for all large n and each $1 \leq k \leq c \log n$,

$$\mathbb{E}\left[m^*_{k,\hat{T}_n}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}) - \hat{m}_{k,\hat{T}_n}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X})\right]^2 \leq n^{-\eta},$$

where the sample tree growing rule $\hat{T}_n$ is defined in Definition 1.24.6 (except with no set of sample indices specified because we are using the special forest model with $\#\tilde{\mathcal{A}} = 1$), the population tree model m^* is defined in (1.64) in Definition 1.24.11, and the tree model $\hat{m}_{k,\hat{T}_n}$ is defined in (1.61).

Given one sample tree growing rule $\hat{T}_n$, this lemma establishes the consistency rate at which the sample tree model approaches the population tree model with the same rule.

Lemma 1.24.3.2 (Approximation error; Lemma 3 in Chi et al. [2020])**.** Consider a sequence of regression functions $\{m_n\}$ such that m_n satisfies Condition 1 with α_{1n} and q_1. (For notational simplicity, the subscript of m_n is dropped whenever there is no confusion.) That is, assume that Conditions 1, 2, 4, and 5 hold with $\alpha_{1n} \geq 1$ and $q_1 \geq 1$, $0 < \eta < 1/8$, $0 < c < \eta/\log 2$, $0 < \rho < 1$, and $\mathbb{E}|\epsilon_1|^q < \infty$ for a sufficiently large q. Then it holds that for all large n and $k = \lfloor c \log n \rfloor$,

$$\mathbb{E}\left[m(\boldsymbol{X}) - m^*_{k,\hat{T}_n}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X})\right]^2 \leq \alpha_{1n} 2^{1/q_1} (\log n)^{-\frac{1-\rho}{q_1}}.$$

Moreover, for the regular case with $q_1 = 1$ in Condition 1, it holds that for all large n and $k = \lfloor c \log n \rfloor$,

$$\mathbb{E}\left[m(\boldsymbol{X}) - m^*_{k,\hat{T}_n}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X})\right]^2 \leq 2\alpha_{1n} n^{-\eta} + \left(\frac{M_0^2}{1-\gamma_0}\right) n^{c \log(1-\gamma_0/(2\alpha_{1n}))}. \tag{1.66}$$

Note the similarity of this result to Theorem 1.24.2.2 (Theorem 1 in Chi et al. [2020]). The authors prove this result by showing that the tree growing rule using the sample CART-split criterion (with some minor manipulation) satisfies Condition 3; then they use Theorem 1.24.2.2 to complete the analysis. (To verify that Condition 3 is satisfied, they use high-dimensional estimation theory for sample trees from Theorem 5.)

In the regular case with $q_1 = 1$, if the learning rate is decreasing (i.e. α_{1n} is increasing), the second term of the approximation bound (1.66) dominates. Using the asymptotic expansion

$$\log\left(1 - \frac{\gamma_0}{2\alpha_{1n}}\right) \approx -\frac{\gamma_0}{2\alpha_{1n}},$$

the rate can be simplified as

$$\mathcal{O}\left(n^{-\frac{c\gamma_0}{2\alpha_{1n}}}\right),$$

which means that the approximation error is guaranteed to be controlled when the inverse of impurity decrease rate grows no faster than the order of $\log n$.

Theorem 1.24.3.3 (Consistency rates; Theorem 2 in Chi et al. [2020])**.** Assume that Conditions 1, 2, 4, and 5 hold with $\alpha_{1n} \geq 1$ and $q_1 \geq 1, 0 < \eta < 1/8, 0 < c < \eta/\log(2), 0 < \rho < 1$. and $\mathbb{E}|\epsilon_1|^q < \infty$ for a sufficiently large q. Assume also that for each $\mathcal{A} \in \tilde{\mathcal{A}}$ it holds that $\#\mathcal{A} = \lceil bn \rceil$ for some $b \in (0, 1]$. Then it holds that for all large n and $k = \lfloor c \log \lceil bn \rceil \rfloor$,

$$\mathbb{E}\left[m(\boldsymbol{X}) - \mathbb{E}\left(\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k, \hat{T}_{(n,\mathcal{A})}, \mathcal{A}}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n\right) \mid \boldsymbol{X}, \mathcal{X}_n\right)\right]^2 \leq \alpha_{1n} 2^{2+q_1^{-1}} \left(\log \lceil bn \rceil\right)^{-\frac{1-\rho}{q_1}}.$$

Moreover, for the regular case with $q_1 = 1$ in Condition 1, it holds that for all large n and $k = \lfloor c \log \lceil bn \rceil \rfloor$,

$$\mathbb{E}\left[m(\boldsymbol{X}) - \mathbb{E}\left(\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k, \hat{T}_{(n,\mathcal{A})}, \mathcal{A}}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n\right) \mid \boldsymbol{X}, \mathcal{X}_n\right)\right]^2 \leq 8\alpha_{1n} \left(\log \lceil bn \rceil\right)^{-\eta} + \left(\frac{4M_0^2}{1 - \gamma_0}\right) \left(\lceil bn \rceil\right)^{c \log(1 - \gamma_0/[2\alpha_{1n}])}.$$

This is the first result on the consistency rate for the original version of the random forest algorithm. With the consistency rate, a direct application of Jensen's inequality allows extending the technical analysis without subsampling (bagging) to that with subsampling.

Definition 1.24.12 (Relevant feature; Definition 5 in Chi et al. [2020])**.** A feature j is said to be **relevant** for $m(\cdot)$ if and only if $0 < \mathbb{E}\left[\mathrm{Var}(m \mid X_s, s \in [p] \setminus j)\right] < \infty$. A feature j is said to be **irrelevant** for $m(\cdot)$ if and only if $\mathbb{E}\left[\mathrm{Var}(m \mid X_s, s \in [p] \setminus j)\right] = 0$).

The authors denote by S^* the set of all relevant features. Let $\{m_n(\cdot)\}$ be a given sequence of regression functions. The authors introduce an additional natural regularity condition to characterize the magnitude of relevance for relevant features defined in Definition 5.

Condition 6. There exists some constant $\iota \geq 0$ such that for each $n \geq 1$ and $j \in S_n^*$, $\mathbb{E}\left[\mathrm{Var}(m_n \mid X_s, s \in [p] \setminus \{j\})\right] \geq n^{-\iota}$, where $\#S_n^* = s_n$ for a sequence $\{s_n\}$.

Theorem 1.24.3.4 (Role of relevance; Theorem 3 in Chi et al. [2020])**.** Assume that Conditions 2, 5, and 6 hold and for some $j \in S^*$ it holds that for each $\mathcal{A} \in \tilde{\mathcal{A}}$, $\hat{T}_{(n,\mathcal{A})}$ is such that feature j is not involved in the sample CART-split. Then we have

$$\mathbb{E}\left[m_n(\boldsymbol{X}) - \mathbb{E}\left(\frac{1}{\#\tilde{\mathcal{A}}} \sum_{\mathcal{A} \in \tilde{\mathcal{A}}} \hat{m}_{k, \hat{T}_{(n,\mathcal{A})}, \mathcal{A}}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n\right) \mid \boldsymbol{X}, \mathcal{X}_n\right)\right]^2 \geq n^{-\iota}.$$

This characterizes the precise contribution of each relevant feature toward the consistency of random forest. On the other hand, the consistency result in Theorem 1.24.3.3 can hold. Therefore if parameter ι is appropriately chosen, a feature screening or selection method may be developed on the basis of Condition 6 introduced above.

1.24.4 A General Estimation Foundation (Section 5 of Chi et al. [2020])

Theorem 1.24.4.1 (Conditional mean estimation; Theorem 4 in Chi et al. [2020])**.** Assume that Conditions 2 and 5 hold, $\mathbb{E}|\epsilon_1|^q < \infty$ for a sufficiently large $q > 0$, and $0 < \eta < 1/2$. Then there exists some constant $c > 0$ such that for all large n and each $1 \leq k \leq c \log n$,

$$\mathbb{E}\left(\sup_T \mathbb{E}\left[m^*_{k,T^\#}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}\right) - \hat{m}^*_{k,T^\#}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n\right)^2 \mid \boldsymbol{\Theta}, \mathcal{X}_n\right]\right) \leq n^{-\eta},$$

where the supremum is over all possible deterministic tree growing rules.

Theorem 5 below characterizes the quality of the sample tree growing rule by showing that with mild adjustment it satisfies Condition 3. Then its quality is justified by Theorem 1.24.2.2.

Theorem 1.24.4.2 (Sample tree estimation; Theorem 4 in Chi et al. [2020])**.** Assume that Conditions 2 and 5 hold and $\mathbb{E}(|\epsilon_1|^q) < \infty$ for a sufficiently large q. Let $0 < \eta < 1/8$, $c > 0$, and δ with $2\eta < \delta < 1/4$ be given. Then there exists an $\mathcal{X}_n$-measurable event $\boldsymbol{U}_n$ such that for all large n, each $1 \leq k \leq c \log n$, each $\epsilon \geq n^{-\eta}$, and $\alpha_2 = 2$, we have that conditional on event $\boldsymbol{U}_n$,

$$\hat{T}_{n,k,n^{-\delta}} \text{ satisfies Condition 3 with } k, \epsilon, \alpha_2.$$

Moreover, for all large n it holds that $\mathbb{P}(\boldsymbol{U}_n^c) \leq n^{-1}$.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Tsahkadsor, Armenia, USSR, 1973.

A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001. ISSN 0162-1459. doi: 10.1198/016214501753208942. URL https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

D. Bertsekas. *Dynamic Programming and Optimal Control*. Number v. 1 in Athena Scientific optimization and computation series. Athena Scientific, 2012a. ISBN 9781886529434. URL https://books.google.com/books?id=REp7swEACAAJ.

D. Bertsekas. *Dynamic Programming and Optimal Control*. Number v. 2 in Athena Scientific optimization and computation series. Athena Scientific, 2012b. ISBN 9781886529441. URL https://books.google.com/books?id=H-PSMwEACAAJ.

L. Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384, 1995. URL https://www-jstor-org.libproxy2.usc.edu/stable/pdf/1269730.pdf?refreqid=excelsior{%}3A76eea9bd08301e990d7d6edd86067262.

O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'institut Henri Poincare (B) Probability and Statistics*, 48(4):1148–1185, 2012. ISSN 02460203. doi: 10.1214/11-AIHP454. URL www.imstat.org/aihpAnnalesdel'.

C.-M. Chi, P. Vossler, Y. Fan, and J. Lv. Asymptotic Properties of High-Dimensional Random Forests *. Technical report, 2020.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian Mean Estimators. *Annals of Statistics*, 44(6):2695–2725, 2016. ISSN 00905364. doi: 10.1214/16-AOS1440.

D. L. Donoho and I. M. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81 (3):425–455, 1994. URL https://www-jstor-org.libproxy2.usc.edu/stable/pdf/2337118.pdf?refreqid=excelsior{%}3Afb36dc2b9ad4d3d57225eebb75e05df2.

G. M. James, P. Radchenko, and J. Lv. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(1):127–142, 2009. URL https://rss-onlinelibrary-wiley-com.libproxy1.usc.edu/doi/pdf/10.1111/j.1467-9868.2008.00668.x.

T. Mathieu and S. Minsker. Excess risk bounds in robust empirical risk minimization. 2019. URL https://arxiv.org/abs/1910.07485.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. ISSN 1537-2715. doi: 10.1080/10618600.2000.10474883. URL https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. URL https://www.andrew.cmu.edu/user/kk3n/simplicity/schwarzbic.pdf.

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.