E-companion

EC.1. Alternative Approaches using Empirical Risk Minimization

In the beginning of Section 2, we noted that the empirical distribution is insufficient for approximating the full-information problem (2). The solution was to consider local neighborhoods in approximating conditional expected costs; these were computed separately for each x. Another approach would be to develop an explicit decision rule and impose structure on it. In this section, we consider an approach to constructing a predictive prescription by selecting from a family of linear functions restricted in some norm,

$$\mathcal{F} = \left\{ z(x) = Wx \,:\, W \in \mathbb{R}^{d_z \times d_x}, \, ||W|| \leq R \right\}, \tag{EC.1}$$

so to minimize the empirical marginal expected costs as in (4),

$$\hat{z}_N(\cdot) \in \arg\min_{z(\cdot)\in\mathcal{F}} \tfrac{1}{N} \textstyle\sum_{i=1}^N c(z(x^i); y^i).$$

The linear decision rule can be generalized by transforming X to include nonlinear terms or by embedding in a reproducing kernel Hilbert space. We consider two examples of a norm on the matrix of linear coefficients, W: the row-wise p,p'-norm and the Schatten p-norm, which are, respectively,

$$||W|| = ||(\gamma_1 \|W_1\|_p, \ldots, \gamma_d \|W_d\|_p)||_{p'}, \quad ||W|| = \big|\big|(\tau_1, \ldots, \tau_{\min\{d_z,d_x\}})\big|\big|_p,$$

where τ_i are W's singular values. For example, the Schatten 1-norm is the matrix nuclear norm. In either case, the restriction on the norm is equivalent to an appropriately-weighted regularization term incorporated into the objectives of (4).

Problem (4) corresponds to the traditional framework of empirical risk minimization in statistical learning with a general loss function. It is also closely related to M-estimation Geer (2000), except that we are concerned with out-of-sample performance rather than inference, an infinite-dimensional decision rule rather than a finite-dimensional parameter, and potentially non-smooth functions. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z;y) = (z-y)^2$, problem (4) corresponds to least-squares regression. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z;y) = (y-z)(\tau - \mathbb{I}\,[y-z<0])$, problem (4) corresponds to quantile regression (cf. Koenker (2005)), which estimates the conditional τ-quantile as a function of x. Rearranging terms, $c(z;y) = (y-z)(\tau - \mathbb{I}\,[y-z<0]) = \max\{(1-\tau)(z-y), \tau(y-z)\}$ is the same as the newsvendor cost function where τ is the service level requirement as observed by Rudin and Vahn (2015). Standard ERM generalization theory deals only with univariate-valued functions. Because most OR/MS problems involve multivariate uncertainty and decisions, in this section we generalize the approach and its associated theoretical guarantees to such multivariate

problems ($d_y \geq 1$, $d_z \geq 1$). In particular, we generalize Rademacher complexity to multivariate-valued decision rules and extend the Rademacher comparison Lemma (Theorem 4.12 of Ledoux and Talagrand (1991)) to this new definition. We can then apply standard results to obtain out-of-sample guarantees.

Before continuing, we note a few limitations of any approach based on (4). For general problems, there is no reason to expect that optimal solutions will have a linear structure (whereas certain distributional assumptions lead to such conclusions in least-squares and quantile regression analyses). In particular, unlike the predictive prescriptions studied in Section 2, the approach based on (4) does not enjoy the same universal guarantees of asymptotic optimality. Instead, we will only have out-of-sample guarantees that depend on our class $\mathcal{F}$ of decision rules.

Another limitation is the difficulty in restricting the decisions to a constrained feasible set $\mathcal{Z} \neq \mathbb{R}^{d_z}$. Consider, for example, the portfolio allocation problem from Section 1.1, where we must have $\sum_{i=1}^{d_x} z_i = 1$. One approach to applying (4) to this problem might be to set $c(z;y) = \infty$ for $z \notin \mathcal{Z}$ (or, equivalently, constrain $z(x^i) \in \mathcal{Z}$ $\forall i$). However, not only will this not guarantee that $z(x) \in \mathcal{Z}$ for x outside the dataset, but we would also run into a problem of infeasibility as we would have N linear equality constraints on $d_z \times d_x$ linear coefficients (a constraint such as $\sum_{i=1}^{d_x} z_i \leq 1$ that does not reduce the affine dimension will still lead to an undesirably flat linear decision rule as N grows). Another approach may be to compose $\mathcal{F}$ with a projection onto $\mathcal{Z}$, but this will generally lead to a non-convex optimization problem that is intractable to solve. Therefore, the approach is limited in its applicability to OR/MS problems.

In a few limited cases, we may be able to sensibly extend the cost function synthetically outside the feasible region while maintaining convexity. For example, in the shipment planning example of Section 1.1, we may allow negative order quantities z and extend the first-stage costs to depend only on the positive part of z, i.e. $p_1 \sum_{i=1}^{d_z} \max\{z_i, 0\}$ (but leave the second-stage costs as they are for convexity). Now, if after training $\hat{z}_N(\cdot)$, we transform any resulting decision by only taking the positive part of each order quantity, we end up with a feasible decision rule whose costs are no worse than the synthetic costs of the original rule.

In the rest of this section we consider the application of the approach (4) to problems where y and z are multivariate and $c(z;y)$ is general, but only treat unconstrained decisions $\mathcal{Z} = \mathbb{R}^{d_z}$.

EC.1.1. Tractability

We first develop sufficient conditions for the problem (4) to be optimized in polynomial time. The proof appears in Section EC.4.

THEOREM EC.1. *Suppose that $c(z;y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z. Then for any fixed x we can find an ϵ-optimal solution to* (4) *in time and oracle calls polynomial in n, d, $\log(1/\epsilon)$ for $\mathcal{F}$ as in* (EC.1).

EC.1.2. Out-of-Sample Guarantees

Next, we characterize the out-of-sample guarantees of a predictive prescription derived from (4). All proofs are in the E-companion. In the traditional framework of empirical risk minimization in statistical learning such guarantees are often derived using Rademacher complexity but these only apply to univariate problems (c.f. Bartlett and Mendelson (2003)). Because most OR/MS problems are multivariate, we generalize this theory appropriately. We begin by generalizing the definition of Rademacher complexity to multivariate-valued functions.

DEFINITION EC.1. Given a sample $S_N = \{s_1, \ldots, s_N\}$, The *empirical multivariate Rademacher complexity* of a class of functions $\mathcal{F}$ taking values in $\mathbb{R}^d$ is defined as

$$\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N) = \mathbb{E}\left[\tfrac{2}{N}\sup_{g\in\mathcal{F}}\sum_{i=1}^n\sum_{k=1}^d \sigma_{ik} g_k(s_i)\middle| s_1, \ldots, s_n\right]$$

where σ_{ik} are independently equiprobably $+1, -1$. The *marginal multivariate Rademacher complexity* is defined as the expectation over the sampling distribution of S_N: $\mathfrak{R}_N(\mathcal{F}) = \mathbb{E}\left[\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N)\right]$.

Note that given only data S_N, the quantity $\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N)$ is observable. Note also that when $d = 1$ the above definition coincides with the common definition of Rademacher complexity.

The theorem below relates the multivariate Rademacher complexity of $\mathcal{F}$ to out-of-sample guarantees on the performance of the corresponding predictive prescription $\hat{z}_N(x)$ from (4). A generalization of the following to mixing processes is given in the supplemental Section EC.3. We denote by $S_N^x = \{x^1, \ldots, x^N\}$ the restriction of our sample to data on X.

THEOREM EC.2. *Suppose $c(z; y)$ is bounded and equi-Lipschitz in z:*

$$\sup_{z\in\mathcal{Z},\, y\in\mathcal{Y}} c(z; y) \le \overline{c}, \quad \sup_{z\neq z'\in\mathcal{Z},\, y\in\mathcal{Y}} \tfrac{c(z;y)-c(z';y)}{||z_k - z'_k||_\infty} \le L < \infty.$$

Then, for any $\delta > 0$, each of the following events occurs with probability at least $1-\delta$,

$$\mathbb{E}\left[c(z(X); Y)\right] \le \tfrac{1}{N}\sum_{i=1}^N c(z(x^i); y^i) + \overline{c}\sqrt{\log(1/\delta')/2N} + L\mathfrak{R}_N(\mathcal{F}) \qquad \forall z \in \mathcal{F}, \qquad \text{(EC.2)}$$

$$\mathbb{E}\left[c(z(X); Y)\right] \le \tfrac{1}{N}\sum_{i=1}^N c(z(x^i); y^i) + 3\overline{c}\sqrt{\log(2/\delta'')/2N} + L\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N^x) \qquad \forall z \in \mathcal{F}. \qquad \text{(EC.3)}$$

In particular, these hold for $z = \hat{z}_N(\cdot) \in \mathcal{F}$.

Equations (EC.2) and (EC.3) provide a bound on the out-of-sample performance of any predictive prescription $z(\cdot) \in \mathcal{F}$. The bound is exactly what we minimize in problem (4) because the extra terms do not depend on $z(\cdot)$. That is, we minimize the empirical risk, which, with additional confidence terms, bounds the true out-of-sample costs of the resulting predictive prescription $\hat{z}_N(\cdot)$.

To prove Theorem EC.2, we first establish a comparison lemma that is an extension of Theorem 4.12 of Ledoux and Talagrand (1991) to our multivariate case.

LEMMA EC.1. *Suppose that c is L-Lipschitz uniformly over y with respect to ∞-norm:*

$$\sup_{z\neq z'\in\mathcal{Z},\, y\in\mathcal{Y}} \frac{c(z;y)-c(z';y)}{\max_{k=1,\ldots,d}|z_k-z'_k|} \leq L < \infty.$$

Let $\mathcal{G}=\{(x,y)\mapsto c\left(f(x);y\right) : f\in\mathcal{F}\}$. Then we have that $\widehat{\mathfrak{R}}_n(\mathcal{G};S_N)\leq L\widehat{\mathfrak{R}}_n(\mathcal{F};S_N^x)$ and therefore also that $\mathfrak{R}_n(\mathcal{G})\leq L\mathfrak{R}_n(\mathcal{F})$. (Notice that one is a univariate complexity and one multivariate and that the complexity of $\mathcal{F}$ involves only the sampling of x.)

Proof Write $\phi_i(z)=c(z;y^i)/L$. Then by Lipschitz assumption and by part 2 of Proposition 2.2.1 from Bertsekas et al. (2003), for each i, ϕ_i is 1-Lipchitz. We now would like to show the inequality in

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\mathcal{G};S_N) &= \mathbb{E}\left[\frac{2}{n}\sup_{z\in\mathcal{F}}\sum_{i=1}^n \sigma_{i0}\phi_i(z(x^i)) \,\middle|\, S_N\right] \\
&\leq L\mathbb{E}\left[\frac{2}{n}\sup_{z\in\mathcal{F}}\sum_{i=1}^n\sum_{k=1}^d \sigma_{ik}z_k(x^i) \,\middle|\, S_N^x\right] \\
&= L\widehat{\mathfrak{R}}_n(\mathcal{F};S_N^x).
\end{aligned}$$

By conditioning and iterating, it suffices to show that for any $T\subset\mathbb{R}\times\mathcal{Z}$ and 1-Lipchitz ϕ,

$$\mathbb{E}\left[\sup_{t,z\in T}\left(t+\sigma_0\phi(z)\right)\right] \leq \mathbb{E}\left[\sup_{t,z\in T}\left(t+\sum_{k=1}^d\sigma_k z_k\right)\right]. \tag{EC.4}$$

The expectation on the left-hand-side is over two values ($\sigma_0=\pm1$) so there are two choices of (t,z), one for each scenario. Let any $(t^{(+1)},z^{(+1)}),(t^{(-1)},z^{(-1)})\in T$ be given. Let k^* and $s^*=\pm1$ be such that

$$\max_{k=1,\ldots,d}\left|z_k^{(+1)}-z_k^{(-1)}\right| = s^*\left(z_{k^*}^{(+1)}-z_{k^*}^{(-1)}\right).$$

Fix $(\tilde{t}^{(\pm1)},\tilde{z}^{(\pm1)})=(t^{(\pm s^*)},z^{(\pm s^*)})$. Then, since these are feasible choices in the inner supremum, choosing $(t,z)(\sigma)=(\tilde{t}^{(\sigma_{k^*})},\tilde{z}^{(\sigma_{k^*})})$, we see that the right-hand-side of (EC.4) has

$$\begin{aligned}
\text{RHS (EC.4)} &\geq \frac{1}{2}\mathbb{E}\left[\tilde{t}^{(+1)}+\tilde{z}_{k^*}^{(+1)}+\sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(+1)}\right] \\
&\quad+\frac{1}{2}\mathbb{E}\left[\tilde{t}^{(-1)}-\tilde{z}_{k^*}^{(-1)}+\sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(-1)}\right] \\
&= \frac{1}{2}\left(t^{(+1)}+t^{(-1)}+\max_{k=1,\ldots,d}\left|z_k^{(+1)}-z_k^{(-1)}\right|\right) \\
&\geq \frac{1}{2}\left(t^{(+1)}+\phi\left(z^{(+1)}\right)\right)+\frac{1}{2}\left(t^{(-1)}-\phi\left(z^{(-1)}\right)\right)
\end{aligned}$$

where the last inequality is due to the Lipschitz condition. Since true for any $(t^{(\pm1)},z^{(\pm1)})$ given, taking suprema over the left-hand-side completes the proof. □

Next, we restate the main result of Bartlett and Mendelson (2003):

THEOREM EC.3. *Consider a class $\mathcal{G}$ of functions $\mathcal{U} \to \mathbb{R}$ that are bounded: $|g(u)| \le \overline{g}$ $\forall g \in \mathcal{G}$, $u \in \mathcal{U}$. Consider a sample $S_n = (u^1, \ldots, u^N)$ of some random variable $T \in \mathcal{T}$. Fix $\delta > 0$. Then we have that, with probability $1-\delta$,*

$$\mathbb{E}\left[g(T)\right] \le \frac{1}{N}\sum_{i=1}^{N} g(u^i) + \overline{g}\sqrt{\log(1/\delta)/2N} + \mathfrak{R}_N(\mathcal{G}) \qquad \forall g \in \mathcal{G}, \tag{EC.5}$$

and that, again with probability $1-\delta$,

$$\mathbb{E}\left[g(T)\right] \le \frac{1}{N}\sum_{i=1}^{N} g(u^i) + 3\overline{g}\sqrt{\log(2/\delta)/2N} + \widehat{\mathfrak{R}}_N(\mathcal{G}) \qquad \forall g \in \mathcal{G}. \tag{EC.6}$$

Finally, we can prove Theorem EC.2:

Proof of Theorem EC.2 Apply Theorem EC.3 to the random variable $U = (X, Y)$ and function class $\mathcal{G} = \{(x,y) \mapsto c\left(f(x); y\right) : f \in \mathcal{F}\}$. Note that by assumption we have boundedness of functions in $\mathcal{G}$ by the constant $\overline{c}$. Bound the complexity of $\mathcal{G}$ by that of $\mathcal{F}$ using Lemma EC.1 and the assumption of $c(z;y)$ being L-Lipschitz. Equations (EC.8) and (EC.9) hold for every $g \in \mathcal{G}$ and hence for every $f \in \mathcal{F}$ and $g(x,y) = c\left(f(x);y\right)$, of which the expectation is the expected costs of the decision rule f. □

Equations (EC.2) and (EC.3) in Theorem EC.2 involve the multivariate Rademacher complexity of our class $\mathcal{F}$ of decision rules. In the next lemmas, we compute appropriate bounds on the complexity of our examples of classes $\mathcal{F}$. The theory, however, applies beyond linear rules.

LEMMA EC.2. *Consider $\mathcal{F}$ as in* (EC.1) *with row-wise p,p' norm for $p \in [2,\infty)$ and $p' \in [1,\infty]$. Let q be the conjugate exponent of p ($1/p + 1/q = 1$) and suppose that $||x||_q \le M$ for all $x \in \mathcal{X}$. Then*

$$\mathfrak{R}_N(\mathcal{F}) \le 2MR\sqrt{\tfrac{p-1}{N}}\textstyle\sum_{k=1}^{d_z}\frac{1}{\gamma_k}.$$

LEMMA EC.3. *Consider $\mathcal{F}$ as in* (EC.1) *with Schatten p-norm. Let $r = \max\{1-1/p, 1/2\}$. Then*

$$\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N^x) \le 2Rd_z^r\sqrt{\tfrac{1}{N}}\sqrt{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\|x^i\|}, \quad \mathfrak{R}_N(\mathcal{F}) \le 2Rd_z^r\sqrt{\tfrac{1}{N}}\sqrt{\mathbb{E}\,||X||_2^2}.$$

The above results indicate that the confidence terms in equations (EC.2) and (EC.3) shrink to 0 as $N \to \infty$ even if we slowly relax norm restrictions. Hence, we can approach the optimal out-of-sample performance over the class $\mathcal{F}$ without restrictions on norms.

Proof of Lemma EC.2 Consider $\mathcal{F}_k = \{z_k(\cdot) : z \in \mathcal{F}\} = \{z_k(x) = w^T x : ||w||_p \le \frac{R}{\gamma_k}\}$, the projection of $\mathcal{F}$ onto the k^{th} coordinate. Then $\mathcal{F} \subset \mathcal{F}_1 \times \cdots \times \mathcal{F}_{d_z}$ and $\mathfrak{R}_N(\mathcal{F}) \le \sum_{k=1}^{d_z}\mathfrak{R}_N(\mathcal{F}_k)$. The latter right-hand-side complexities are the common univariate Rademacher complexities. Applying Theorem 1 of Kakade et al. (2008) to each component we get $\mathfrak{R}_N(\mathcal{F}_k) \le 2M\sqrt{\frac{p-1}{N}}\frac{R}{\gamma_k}$. □

Proof of Lemma EC.3 Let q be p's conjugate exponent ($1/p+1/q=1$). In terms of vector norms on $v\in\mathbb{R}^d$, if $q\geq 2$ then $||v||_p \leq ||v||_2$ and if $q\leq 2$ then $||b||_p \leq d^{1/2-1/p}||v||_2$. Let F be the matrix $F_{ji}=x_j^i$. Note that $F\sigma\in\mathbb{R}^{d_x\times d_z}$. By Jensen's inequality and since Schatten norms are vector norms on singular values,

$$\begin{aligned}
\widehat{\mathfrak{R}}_N^2(\mathcal{F};S_N^x) &\leq \frac{4}{N^2}\mathbb{E}\left[\sup_{||W||_p\leq R}\operatorname{Trace}\left(WF\sigma\right)^2\middle|S_N^x\right]\\
&= \frac{4R^2}{N^2}\mathbb{E}\left[||F\sigma||_q^2\middle|S_N^x\right]\\
&\leq \frac{4R^2}{N^2}\max\left\{\min\left\{d_z,d_x\right\}^{1-2/p},1\right\}\mathbb{E}\left[||F\sigma||_2^2\middle|S_N^x\right]\\
&\leq \frac{4R^2}{N^2}\max\left\{d_z^{1-2/p},1\right\}\mathbb{E}\left[||F\sigma||_2^2\middle|S_N^x\right].
\end{aligned}$$

The first result follows because

$$\begin{aligned}
\frac{1}{N}\mathbb{E}\left[||F\sigma||_2^2\,\middle|S_N^x\right] &= \frac{1}{n}\sum_{k=1}^{d_z}\sum_{j=1}^{d_x}\sum_{i,i'=1}^{N}x_j^i x_j^{i'}\mathbb{E}\left[\sigma_{ik}\sigma_{i'k}\right]\\
&= \frac{d_z}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_x}(x_j^i)^2 = d_z\,\widehat{\mathbb{E}}_N\,||x||_2^2
\end{aligned}$$

The second result follows by applying Jensen's inequality again to pass the expectation over S_n into the square. □

EC.2. Extensions of Asymptotic Optimality to Mixing Processes and Proofs

In this supplemental section, we generalize the asymptotic results to mixing process and provide the omitted proofs from Sections 4.3 and 4.4.

EC.2.1. Mixing Processes

We begin by defining stationary and mixing processes.

Definition EC.2. A sequence of random variables $V_1, V_2, \ldots$ with joint measure μ is called *stationary* if joint distributions of finitely many consecutive variables are invariant to shifting. That is,

$$\mu_{V_t,\ldots,V_{t+k}} = \mu_{V_s,\ldots,V_{s+k}} \quad \forall s,t\in\mathbb{N},\, k\geq 0,$$

where $\mu_{V_t,\ldots,V_{t+k}}$ is the induced measure on a sequence of length k.

In particular, if a sequence is stationary then the variables have identical marginal distributions, but they may not be independent and the sequence may not be exchangeable. Instead of independence, mixing is the property that if standing at particular point in the sequence we look far enough ahead, the head and the tail look nearly independent, where "nearly" is defined by different metrics for different definitions of mixing.

DEFINITION EC.3. Given a stationary sequence $\{V_t\}_{t\in\mathbb{N}}$, denote by $\mathcal{A}^t = \sigma(V_1,\dots,V_t)$ the sigma-algebra generated by the first t variables and by $\mathcal{A}_t = \sigma(V_t, V_{t+1},\dots)$ the sigma-algebra generated by the subsequence starting at t. Define the *mixing coefficients at lag k*

$$\begin{aligned}
\alpha(k) &= \sup_{t\in\mathbb{N},\, A\in\mathcal{A}^t,\, B\in\mathcal{A}_{t+k}} |\mu(A\cap B) - \mu(A)\mu(B)| \\
\beta(k) &= \sup_{t\in\mathbb{N}} \left|\left| \mu_{\{V_s\}_{s\le t}} \otimes \mu_{\{V_s\}_{s\ge t+k}} - \mu_{\{V_s\}_{s\le t \vee s\ge t+k}} \right|\right|_{\mathrm{TV}} \\
\rho(k) &= \sup_{t\in\mathbb{N},\, Q\in L_2(\mathcal{A}^t),\, R\in L_2(\mathcal{A}_{t+k})} |\mathrm{Corr}(Q,R)|
\end{aligned}$$

where $||\cdot||_{\mathrm{TV}}$ is the total variance and $L_2(\mathcal{A})$ is the set of $\mathcal{A}$-measurable square-integrable real-valued random variables.

$\{V_t\}$ is said to be α-mixing if $\alpha(k) \overset{k\to\infty}{\longrightarrow} 0$, β-mixing if $\beta(k) \overset{k\to\infty}{\longrightarrow} 0$, and ρ-mixing if $\rho(k) \overset{k\to\infty}{\longrightarrow} 0$.

Notice that an iid sequence has $\alpha(k) = \beta(k) = \rho(k) = 0$. Bradley (1986) establishes that $2\alpha(k) \le \beta(k)$ and $4\alpha(k) \le \rho(k)$ so that either β- or ρ-mixing implies α-mixing.

Many processes satisfy mixing conditions under mild assumptions: auto-egressive moving-average (ARMA) processes (cf. Mokkadem (1988)), generalized autoregressive conditional heteroskedasticity (GARCH) processes (cf. Carrasco and Chen (2002)), and certain Markov chains. For a thorough discussion and more examples see Doukhan (1994) and Bradley (2005). Mixing rates are often given explicitly by model parameters but they can also be estimated from data (cf. Mcdonald et al. (2011)). Sampling from such processes models many real-life sampling situations where observations are taken from an evolving system such as, for example, the stock market, inter-dependent product demands, or aggregates of doubly stochastic arrival processes as in the posts on social media.

EC.2.2. Asymptotic Optimality

Let us now restate the results of Section 4.3 in more general terms, encompassing both iid and mixing conditions on S_N.

THEOREM EC.4 **(kNN)**. *Suppose Assumptions 3, 4, and 5 hold and that S_N is generated by iid sampling. Let $w_{N,i}(x)$ be as in (12) with $k = \min\left\{\lceil CN^\delta\rceil, N-1\right\}$ for some $C>0$, $0<\delta<1$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

THEOREM EC.5 **(Kernel Methods)**. *Suppose Assumptions 3, 4, and 5 hold and that $\mathbb{E}\left[|c(z;Y)|\max\left\{\log|c(z;Y)|,0\right\}\right] < \infty$ for each z. Let $w_{N,i}(x)$ be as in (13) with K being any of the kernels in Section 2.2 and $h = CN^{-\delta}$ for $C,\delta>0$. Let $\hat{z}_N(x)$ be as in (3). If S_N comes from*

1. *an iid process and $\delta < 1/d_x$, or*
2. *a ρ-mixing process with $\rho(k) = O(k^{-\gamma})$ $(\gamma>0)$ and $\delta < 2\gamma/(d_x + 2d_x\gamma)$, or*
3. *an α-mixing process with $\alpha(k) = O(k^{-\gamma})$ $(\gamma>1)$ and $\delta < 2(\gamma-1)/(3d_x + 2d_x\gamma)$,*

then $\hat{z}_N(x)$ is asymptotically optimal and consistent.

THEOREM EC.6 **(Recursive Kernel Methods)**. *Suppose Assumptions 3, 4, and 5 hold and that S_N comes from a ρ-mixing process with $\sum_{k=1}^{\infty}\rho(k)<\infty$ (or iid). Let $w_{N,i}(x)$ be as in (14) with K being the naïve kernel and with $h_i = Ci^{-\delta}$ for some $C>0$, $0<\delta<1/(2d_x)$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

THEOREM EC.7 **(Local Linear Methods)**. *Suppose Assumptions 3, 4, and 5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X and twice continuously differentiable, and that costs are bounded over y for each z (i.e., $|c(z;y)|\le g(z)$) and twice continuously differentiable. Let $w_{N,i}(x)$ be as in (15) with K being any of the kernels in Section 2.2 and with $h_N = CN^{-\delta}$ for some $C,\delta>0$. Let $\hat{z}_N(x)$ be as in (3). If S_N comes from*

1. *an iid process and $\delta<1/d_x$, or*
2. *an α-mixing process with $\alpha(k)=O(k^{-\gamma})$, $\gamma>d_x+3$, and $\delta<(\gamma-d_x-3)/(d_x(\gamma-d_x+3))$,*

then $\hat{z}_N(x)$ is asymptotically optimal and consistent.

THEOREM EC.8 **(Nonnegative Local Linear Methods)**. *Suppose Assumptions 3, 4, and 5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X and twice continuously differentiable, and that costs are bounded over y for each z (i.e., $|c(z;y)|\le g(z)$) and twice continuously differentiable. Let $w_{N,i}(x)$ be as in (16) with K being any of the kernels in Section 2.2 and with $h_N = CN^{-\delta}$ for some $C,\delta>0$. Let $\hat{z}_N(x)$ be as in (3). If S_N comes from*

1. *an iid process and $\delta<1/d_x$, or*
2. *an α-mixing process with $\alpha(k)=O(k^{-\gamma})$, $\gamma>d_x+3$, and $\delta<(\gamma-d_x-3)/(d_x(\gamma-d_x+3))$,*

then $\hat{z}_N(x)$ is asymptotically optimal and consistent.

THEOREM EC.9 **(Decisions Affect Uncertainty)**. *Suppose Assumptions 1, 2, 3, 4, and 5 (case 1) hold, that $\mu_{(X,Z_1)}$ is absolutely continuous and has density bounded away from 0 and ∞ on the support of X,Z_1 and twice continuously differentiable, and that costs are bounded over y for each z (i.e., $|c(z;y)|\le g(z)$) and twice continuously differentiable. Let $w_{N,i}(\tilde{x})$ be as in (13), (15), or (16) applied to $\tilde{S}_N$ with K being any of the kernels in Section 2.2 and with $h_N = CN^{-\delta}$ for some $C>0$, $\delta>0$. Let $\hat{z}_N(x)$ be as in (21). If $\tilde{S}_N$ comes from*

1. *an iid process and $\delta<1/(d_x+d_{z_1})$, or*
2. *an α-mixing process with $\alpha(k)=O(k^{-\gamma})$, $\gamma>d_x+d_{z_1}+3$, and $\delta<(\gamma-d_x-d_{z_1}-3)/(d_x(\gamma-d_x-d_{z_1}+3))$,*

Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.

EC.2.3. Proofs of Asymptotic Results for Local Predictive Prescriptions

First, we establish some preliminary results. In what follows, let

$$\begin{aligned}
C(z|x) &= \mathbb{E}\left[c(z;Y)\middle|X=x\right],\\
\widehat{C}_N(z|x) &= \sum_{i=1}^N w_{N,i}(x)c(z;y^i),\\
\mu_{Y|x}(A) &= \mathbb{E}\left[\mathbb{I}\left[Y\in A\right]\middle|X=x\right],\\
\hat{\mu}_{Y|x,N}(A) &= \sum_{i=1}^N w_{N,i}(x)\mathbb{I}\left[y^i\in A\right].
\end{aligned}$$

LEMMA EC.4. *If $\{(x^i,y^i)\}_{i\in\mathbb{N}}$ is stationary and $f:\mathbb{R}^{m_Y}\to\mathbb{R}$ is measurable then $\{(x^i,f(y^i))\}_{i\in\mathbb{N}}$ is also stationary and has mixing coefficients no larger than those of $\{(x^i,y^i)\}_{i\in\mathbb{N}}$.*

Proof This is simply because a transform can only make the generated sigma-algebra coarser. For a single time point, if f is measurable and $B\in\mathcal{B}(\mathbb{R})$ then by definition $f^{-1}(B)\in\mathcal{B}(\mathbb{R})$ and, therefore, $\{Y^{-1}(f^{-1}(B)):B\in\mathcal{B}(\mathbb{R})\}\subset\{Y^{-1}(B):B\in\mathcal{B}(\mathbb{R}^{m_Y})\}$. Here the transform is applied independently across time so the result holds ($f\times\cdots\times f$ remains measurable). □

LEMMA EC.5. *Suppose Assumptions 3 and 4 hold. Fix $x\in\mathcal{X}$ and a sample path of data such that, for every $z\in\mathcal{Z}$, $\widehat{C}_N(z|x)\to C(z|x)$. Then $\widehat{C}_N(z|x)\to C(z|x)$ uniformly in z over any compact subset of $\mathcal{Z}$.*

Proof Let any convergent sequence $z_N\to z$ and $\epsilon>0$ be given. By equicontinuity and $z_N\to z$, $\exists N_1$ such that $|c(z_N;y)-c(z;y)|\le\epsilon/2\ \forall N\ge N_1$. Then $\left|\widehat{C}_N(z_N|x)-\widehat{C}_N(z|x)\right|\le\mathbb{E}_{\hat{\mu}_{Y|x,N}}|c(z_N;y)-c(z;y)|\le\epsilon/2\ \forall N\ge N_1$. By assumption $\widehat{C}_N(z|x)\to C(z|x)$ and hence $\exists N_2$ such that $\left|\widehat{C}_N(z|x)-C(z|x)\right|\le\epsilon/2$. Therefore, for $N\ge\max\{N_1,N_2\}$,

$$\left|\widehat{C}_N(z_N|x)-C(z|x)\right|\le\left|\widehat{C}_N(z_N|x)-\widehat{C}_N(z|x)\right|+\left|\widehat{C}_N(z|x)-C(z|x)\right|\le\epsilon.$$

Hence $\widehat{C}_N(z_N|x)\to C(z|x)$ for any convergent sequence $z_N\to z$.

Now fix $E\subset\mathcal{Z}$ compact and suppose for contradiction that $\sup_{z\in E}\left|\widehat{C}_N(z|x)-C(z|x)\right|\not\to 0$. Then $\exists\epsilon>0$ and $z_N\in E$ such that $\left|\widehat{C}_N(z_N|x)-C(z_N|x)\right|\ge\epsilon$ infinitely often. Restricting first to a subsequence where this always happens and then using the compactness of E, there exists a convergent subsequence $z_{N_k}\to z\in E$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z_{N_k}|x)\right|\ge\epsilon$ for every k. Then,

$$0<\epsilon\le\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z_{N_k}|x)\right|\le\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z|x)\right|+\left|C(z|x)-C(z_{N_k}|x)\right|.$$

Since $z_{N_k}\to z$, we have shown before that $\exists k_1$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z|x)\right|\le\epsilon/2\ \forall k\ge k_1$. By equicontinuity and $z_{N_k}\to z$, $\exists k_2$ such that $\left|c(z_{N_k};y)-c(z;y)\right|\le\epsilon/4\ \forall k\ge k_2$. Hence, also $\left|C(z|x)-C(z_{N_k}|x)\right|\le\mathbb{E}\left[\left|c(z_{N_k};y)-c(z;y)\right|\middle|X=x\right]\le\epsilon/4\ \forall k\ge k_2$. Considering $k=\max\{k_1,k_2\}$ we get the contradiction that $0<\epsilon\le\epsilon/2$. □

LEMMA EC.6. *Suppose Assumptions 3, 4, and 5 hold. Fix $x \in \mathcal{X}$ and a sample path of data such that $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly and, for every $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to C(z|x)$. Then $\lim_{N\to\infty}\left(\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)\right) = v^*(x)$ and every sequence $z_N \in \arg\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)$ satisfies $\lim_{N\to\infty} C\left(z_N|x\right) = v^*(x)$ and $\lim_{N\to\infty} \inf_{z\in Z^*(x)} ||z - z_N|| = 0$.*

Proof First, we show $\widehat{C}_N(z|x)$ and $C(z|x)$ are continuous and eventually coercive. Let $\epsilon > 0$ be given. By equicontinuity, $\exists \delta > 0$ such that $|c(z;y) - c(z';y)| \le \epsilon \ \forall y \in \mathcal{Y}$ whenever $||z - z'|| \le \delta$. Hence, whenever $||z - z'|| \le \delta$, we have $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \le \mathbb{E}_{\hat{\mu}_{Y|x,N}} |c(z;y) - c(z';y)| \le \epsilon$ and $|C(z|x) - C(z'|x)| \le \mathbb{E}\left[|c(z;y) - c(z';y)| \,\middle|\, X = x\right] \le \epsilon$. This gives continuity. Coerciveness is trivial if $\mathcal{Z}$ is bounded. Suppose it is not. Without loss of generality D_x is compact, otherwise we can take any compact subset of it that has positive probability on it. Then by assumption of weak convergence $\exists N_0$ such that $\hat{\mu}_{Y|x,N}(D_x) \ge \mu_{Y|x}(D_x)/2 > 0$ for all $N \ge N_0$. Now let $z_k \in \mathcal{Z}$ be any sequence such that $||z_k|| \to \infty$. Let $M > 0$ be given. Let $\lambda' = \liminf_{k\to\infty} \inf_{y\notin D_x} c(z_k;y)$ and $\lambda = \max\{\lambda', 0\}$. By assumption $\lambda' > -\infty$. Hence $\exists k_0$ such that $\inf_{y\notin D_x} c(z_k;y) \ge \lambda' \ \forall k \ge k_0$. By D_x-uniform coerciveness and $||z_k|| \to \infty$, $\exists k_1 \ge k_0$ such that $c(z_k;y) \ge (2M - 2\lambda)/\mu_{Y|x}(D_x) \ \forall k \ge k_1$ and $y \in D_x$. Hence, $\forall k \ge k_1$ and $N \ge N_0$,

$$C(z_k|x) \ge \mu_{Y|x}(D_x) \times (2M - 2\lambda)/\mu_{Y|x}(D_x) + (1 - \mu_{Y|x}(D_x))\lambda' \ge 2M - 2\lambda + \lambda \ge M,$$
$$\widehat{C}_N(z_k|x) \ge \hat{\mu}_{Y|x,N}(D_x) \times (2M - 2\lambda)/\hat{\mu}_{Y|x,N}(D_x) + (1 - \hat{\mu}_{Y|x,N}(D_x))\lambda' \ge M - \lambda + \lambda = M,$$

since $\alpha\lambda' \ge \lambda$ if $\alpha \ge 0$. This gives coerciveness eventually. By the usual extreme value theorem (c.f. Bertsekas (1999), pg. 669), $\widehat{\mathcal{Z}}_N(x) = \arg\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)$ and $\mathcal{Z}^*(x) = \arg\min_{z\in\mathcal{Z}} C(z|x)$ exist, are nonempty, and are compact.

Now we show there exists $\mathcal{Z}^*_\infty(x)$ compact such that $\mathcal{Z}^*(x) \subset \mathcal{Z}^*_\infty(x)$ and $\widehat{\mathcal{Z}}_N(x) \subset \mathcal{Z}^*_\infty(x)$ eventually. If $\mathcal{Z}$ is bounded this is trivial. So suppose otherwise (and again, without loss of generality D_x is compact). Fix any $z^* \in \mathcal{Z}^*(x)$. Then by Lemma EC.5 we have $\widehat{C}_N(z^*|x) \to C(z^*|x)$. Since $\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \le \widehat{C}_N(z^*|x)$, we have $\limsup_{N\to\infty} \min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \le C(z^*|x) = \min_{z\in\mathcal{Z}} C(z|x) = v^*$. Now suppose for contradiction no such $\mathcal{Z}^*_\infty(x)$ exists. Then there must be a subsequence $z_{N_k} \in \widehat{\mathcal{Z}}_{N_k}$ such that $\left|\left|z_{N_k}\right|\right| \to \infty$. By D_x-uniform coerciveness and $\left|\left|z_{N_k}\right|\right| \to \infty$, $\exists k_1 \ge k_0$ such that $c(z_{N_k};y) \ge 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x) \ \forall k \ge k_1$ and $y \in D_x$. Hence, $\forall k \ge k_1$ and $N \ge N_0$,

$$\widehat{C}_N(z_{N_k}|x) \ge \hat{\mu}_{Y|x,N}(D_x) \times 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x) + \left(1 - \hat{\mu}_{Y|x,N}(D_x)\right) \ge v^* + 1.$$

This yields a contradiction $v^* + 1 \le v^*$. So $\mathcal{Z}^*_\infty(x)$ exists.

Applying Lemma EC.5,

$$\tau_N = \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| \to 0.$$

The first result follows from

$$\delta_N = \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \le \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| = \tau_N \to 0.$$

Now consider any sequence $z_N \in \widehat{\mathcal{Z}}_N(x)$. The second result follows from

$$\left|C\left(\hat{z}_N|x\right) - \min_{z\in\mathcal{Z}} C(z|x)\right| \le \left|\widehat{C}_N(\hat{z}_N(x)|x) - C\left(\hat{z}_N|x\right)\right| + \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \le \tau_N + \delta_N \to 0.$$

Suppose the third result is false. Then since $\mathcal{Z}^*_\infty(x)$ is compact, there is a convergent subsequence $z_{N_k} \to z'$ such that $\inf_{z\in\mathcal{Z}^*(x)} \left|\left|z_{N_k} - z\right|\right| \ge \eta > 0$ for all k. Since $\inf_{z\in\mathcal{Z}^*(x)} ||z' - z|| \ge \inf_{z\in\mathcal{Z}^*(x)} \left|\left|z_{N_k} - z\right|\right| - \left|\left|z_{N_k} - z'\right|\right| \to \eta > 0$, we have $z' \notin \mathcal{Z}^*(x)$ and hence $\epsilon = C(z'|x) - \min_{z\in\mathcal{Z}} C(z|x) > 0$. By equicontinuity and $z_{N_k} \to z'$, $\exists k_2$ such that $\left|c(z_{N_k}; y) - c(z'; y)\right| \le \epsilon/2 \ \forall y \in \mathcal{Y} \ \forall k \ge k_2$. Then, $\left|C(z_{N_k}|x) - C(z'|x)\right| \le \mathbb{E}\left[\left|c(z_{N_k}; y) - c(z'; y)\right| \middle| X = x\right] \le \epsilon/2 \ \forall k \ge k_2$. Therefore, $\forall k \ge k_2$,

$$\tau_{N_k} + \delta_{N_k} \ge C(z_{N_k}|x) - \min_{z\in\mathcal{Z}} C(z|x) \ge C(z'|x) - \min_{z\in\mathcal{Z}} C(z|x) - \epsilon/2 = \epsilon/2,$$

which, taking limits, is a contradiction, yielding the third result. □

Lemma EC.7. *Suppose $c(z;y)$ is equicontinuous in z. Suppose moreover that for each fixed $z \in \mathcal{Z} \subset \mathbb{R}^d$ we have that $\widehat{C}_N(z|x) \to C(z|x)$ a.s. for μ_X-a.e.x and that for each fixed measurable $D \subset \mathcal{Y}$ we have that $\hat{\mu}_{Y|x,N}(D) \to \mu_{Y|x}(D)$ a.s. for μ_X-a.e.x. Then, a.s. for μ_X-a.e.x, $\widehat{C}_N(z|x) \to C(z|x)$ for all $z \in \mathcal{Z}$ and $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly.*

Proof Since Euclidean space is separable, $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly a.s. for μ_X-a.e.x (c.f. Theorem 11.4.1 of Dudley (2002)). Consider the set $\mathcal{Z}' = \mathcal{Z} \cap \mathbb{Q}^d \cup \{\text{the isolated points of } \mathcal{Z}\}$. Then $\mathcal{Z}'$ is countable and dense in $\mathcal{Z}$. Since $\mathcal{Z}'$ is countable, by continuity of measure, a.s. for μ_X-a.e.x, $\widehat{C}_N(z'|x) \to C(z'|x)$ for all $z' \in \mathcal{Z}'$. Restrict to a sample path and x where this event occurs. Consider any $z \in \mathcal{Z}$ and $\epsilon > 0$. By equicontinuity $\exists \delta > 0$ such that $|c(z;y) - c(z';y)| \le \epsilon/2$ whenever $||z - z'|| \le \delta$. By density there exists such $z' \in \mathcal{Z}'$. Then, $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \le \mathbb{E}_{\hat{\mu}_{Y|x,N}}\left[|c(z;y) - c(z';y)|\right] \le \epsilon/2$ and $|C(z|x) - C(z'|x)| \le \mathbb{E}\left[|c(z;y) - c(z';y)| \middle| X = x\right] \le \epsilon/2$. Therefore, $0 \le \left|\widehat{C}_N(z|x) - C(z|x)\right| \le \left|\widehat{C}_N(z'|x) - C(z'|x)\right| + \epsilon \to \epsilon$. Since true for each ϵ, the result follows for all $z \in \mathcal{Z}$. The choice of particular sample path and x constitute a measure-1 event by assumption. □

Now, we prove the general form of the asymptotic results from Section EC.2.2.

Proof of Theorem EC.4 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 5 of Walk (2010) to Y'. By iid sampling and choice of k, we have that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0,1]$. Therefore applying Theorem 5 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma EC.7 we obtain that assumptions for Lemma EC.6 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem EC.5 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 3 of Walk (2010) to Y'. By assumption in theorem statement, we also have that $\mathbb{E}\{|Y'| \max\{\log|Y'|, 0\}\} < \infty$. Moreover each of the kernels in Section 2.2 can be rewritten $K(x) = H(||x||)$ such that $H(0) > 0$ and $\lim_{t\to\infty} t^{d_X} H(t) \to 0$.

Consider the case of iid sampling. Then our data on (X, Y') is ρ-mixing with $\rho(k) = 0$. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Consider the case of ρ-mixing or α-mixing. By Lemma EC.4, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and $\mathbb{E}\{|Y'| \max\{\log|Y'|, 0\}\} \le 1 < \infty$. Therefore applying Theorem 3 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma EC.7 we obtain that assumptions for Lemma EC.6 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem EC.6 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 4 of Walk (2010) to Y'. Note that the naïve kernel satisfies the necessary conditions.

Since our data on (X, Y) is ρ-mixing by assumption, we have that by Lemma EC.4, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choice of the naïve kernel and h_N, Theorem 4 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability. Therefore applying Theorem 4 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma EC.7 we obtain that assumptions for Lemma EC.6 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem EC.7 Fix $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Set $Y' = c(z;Y)$. By Assumption 3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 11 of Hansen (2008) to Y' and use the notation thereof. Fix the neighborhood of consideration to the point x (i.e., set $c_N = 0$) since uniformity in x is not of interest. All of the kernels in Section 2.2 are bounded above and square integrable and therefore satisfy Assumption 1

of Hansen (2008). Let f be the density of X. By assumption $0<\delta\le f(x)\le B_0<\infty$ for all $x\in\mathcal{X}$. Moreover, our choice of h_N satisfies $h_N\to 0$.

Consider first the iid case. Then we have $\alpha(k)=0=O(k^{-\gamma})$ for $\gamma=\infty$ (β in Hansen (2008)). Combined with boundedness conditions of Y' and f ($|Y'|\le g(z)<\infty$ and $\delta<f<B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $\gamma=\infty$, $s=\infty$ in (17) of Hansen (2008) we get $\theta=1$. Therefore, since $h=O(N^{-1/d_x})$ we have

$$\frac{(\log\log N)^4(\log N)^2}{N^\theta h_N^{d_x}}\to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have that $\widehat{C}_N(z|x)\to\mathbb{E}[Y'|X=x]$ a.s.

Now consider the α-mixing case. If the mixing conditions hold for X,Y then by Lemma EC.4, equal or lower mixing coefficients hold for X,Y'. By letting $s=\infty$ we have $\gamma>d_x+3>2$. Combined with boundedness conditions of Y' and f ($|Y'|\le g(z)<\infty$ and $\delta<f<B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $q=\infty$, $s=\infty$ in (16) and (17) of Hansen (2008) we get $\theta=\frac{\gamma-d_x-3}{\gamma-d_x+3}$. Therefore, since $h_N=O(N^{-\theta/d_x})$ we have

$$\frac{(\log\log N)^4(\log N)^2}{N^\theta h_N^{d_x}}\to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have again that $\widehat{C}_N(z|x)\to\mathbb{E}[Y'|X=x]$ a.s.

Since $x\in\mathcal{X}$ was arbitrary we have convergence for μ_X-a.e.x a.s.

Now fix D measurable. Consider a response variable $Y'=\mathbb{I}[y\in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0,1]$. In addition, by Lemma EC.4, equal or lower mixing coefficients hold for X,Y' as hold for X,Y. Therefore applying Theorem 11 of Hansen (2008) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma EC.7 we obtain that assumptions for Lemma EC.6 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem EC.8 Consider the unnormalized local linear weights, which we rewrite as:

$$\begin{aligned}\tilde{w}_{N,i}(x)&=k_i(x)\left(1-\sum_{j=1}^n k_j(x)(x^j-x)^T\Xi(x)^{-1}(x^i-x)\right)\\&=K\left(\frac{x^i-x}{h_N}\right)\left(1-h_N\hat{\Phi}(x)^T\hat{\Psi}(x)^{-1}\left(\frac{x^i-x}{h_N}\right)\right),\\\text{where}\quad\hat{\Phi}(x)&=\frac{1}{Nh_N^{d+1}}\sum_{j=1}^n K\left(\frac{x^j-x}{h_N}\right)\left(\frac{x^j-x}{h_N}\right),\\\hat{\Psi}(x)&=\frac{1}{Nh_N^{d}}\sum_{j=1}^n K\left(\frac{x^j-x}{h_N}\right)\left(\frac{x^j-x}{h_N}\right)\left(\frac{x^j-x}{h_N}\right)^T.\end{aligned}$$

We will show that $\tilde{w}_{N,i}(x) \geq 0$ eventually as N grows for μ_X-a.e.x, a.s. Then we will have that weights (16) are equal to weights (15) eventually and Theorem EC.8 applies. Let $\Sigma = \int K(u)uu^T du$, $a_n^* = \left(\frac{\log N}{N h_N^D}\right)^{1/2} + h_N^2$, and f_X denote the density of X. Using the properties of the kernel (symmetric and zero outside the unit ball) and the differentiability of f_X (series expandable), to show that

$$\begin{aligned}
\mathbb{E}\left[\hat{\Phi}(x)\right] &= \frac{1}{h_N^{d+1}} \int K\left(\frac{x'-x}{h_N}\right)\left(\frac{x'-x}{h_N}\right) f_X(u)dx' \\
&= \frac{1}{h_N}\int K(u)\, u f_X(x+h_N u)du \\
&= \frac{1}{h_N}\int K(u)\, u(f_X(x) + h_N \nabla f_X(x)^T u + O(h_N^2 ||u||^2))du \\
&= \Sigma \nabla f_X(x) + O(h_N^2), \\
\mathbb{E}\left[\hat{\Psi}(x)\right] &= \frac{1}{h_N^{d}} \int K\left(\frac{x'-x}{h_N}\right)\left(\frac{x'-x}{h_N}\right)\left(\frac{x'-x}{h_N}\right)^T f_X(u)dx' \\
&= \int K(u)\, uu^T f_X(x+h_N u)du \\
&= \int K(u)\, uu^T (f_X(x) + h_N \nabla f_X(x)^T u + O(h_N^2 ||u||^2))du \\
&= f_X(x)\Sigma + O(h_N^2).
\end{aligned}$$

By two invocations of Theorem 2 of Hansen (2008), $\hat{\Phi}(x) = \Sigma \nabla f_X(x) + O(a_n^*)$ and $\hat{\Psi}(x) = \Sigma \nabla + O(a_n^*)$ uniformly in x a.s. Note that $K\left(\frac{x^i - x}{h_N}\right) = 0$ if $\left|\left|\frac{x^i-x}{h_N}\right|\right| > 1$ so to show $\tilde{w}_{N,i}(x) \geq 0$ we can restrict to $\left|\left|\frac{x^i-x}{h_N}\right|\right| \leq 1$. Therefore, we have that $\tilde{w}_{N,i}(x) = k_i(x)(1 - O(h_N))$ where $O(h_N)$ is uniformly in x a.s. so that $w_{N,i}(x) \geq 0$ for all i eventually for μ_X-a.e.x, a.s. □

Proof of Theorem EC.9 Fix any x. Let us redefine

$$C(z|x) = \mathbb{E}\left[c(z;Y(z))\middle|X = x\right], \qquad \widehat{C}_N(z|x) = \sum_{i=1}^N w_{N,i}(x,z_1)c(z;y^i).$$

Next, fix any z. Let $Y' = c(z;Y)$ and note that, by Lemma EC.4, the same mixing coefficients hold for $((X,Z_1),Y')$ as do for (X,Y,Z). In the case of weights given by (13), applying Theorem 9 of Hansen (2008) to $((X,Z_1),Y')$ yields the following uniform convergence over the inputs to the weights (x,z_1): we have that, for some $c_N \to \infty$, almost surely

$$\sup_{||x'||+\left|\left|z_1'\right|\right| \leq c_N} \left| \sum_{i=1}^N w_{N,i}(x',z_1')c(z;y^i) - \mathbb{E}\left[c(z;Y)\middle|X = x', Z_1 = z_1'\right]\right| \to 0. \tag{EC.7}$$

For the case of weights given by (15), applying Theorem 11 of Hansen (2008) yields the same result (EC.7). Finally, in the case of weights given by (16), we repeat the argument in Theorem EC.8 verbatim but replacing x by (x,z_1) everywhere to arrive at the conclusion that the weights are eventually nonnegative, eventually reduce to the case of weights given by (15), and yield the same result (EC.7). A critical observation is that (EC.7) holds for $z_1' \neq z_1$. Now, let us restrict

to the almost sure event that (EC.7) holds simultaneously for all $z \in \mathcal{Z} \cap \mathbb{Q}^{d_z}$, of which there are countably many.

Let $\epsilon > 0$ be given. By equicontinuity, for each $z \in \mathcal{Z} \cap \mathbb{Q}^{d_z}$ there is $\delta_z > 0$ such that $|c(z;y) - c(z';y)| \le \epsilon/4$ for all $y, ||z - z'|| \le \delta_z$. By density of rationals, these balls cover $\mathcal{Z}$. Since $\mathcal{Z}$ is compact, there is a finite collection $\tilde{z}^1, \dots, \tilde{z}^k \in \mathcal{Z} \cap \mathbb{Q}^{d_z}$ such that for all $z \in \mathcal{Z}$ there is a j such that $||z - \tilde{z}^j|| \le \delta_{\tilde{z}^j}$. Let N_1 be large enough so that for all $N \ge N_1$, $||x|| + \sup_{z \in \mathcal{Z}} ||z_1|| \le c_N$ and the left-hand-side of (EC.7) is no more than $\epsilon/2$ for each of the finitely-many $\tilde{z}^1, \dots, \tilde{z}^k$.

Let $N \ge N_1$. Let any z be given. Let j be such that $||z - \tilde{z}^j|| \le \delta_{\tilde{z}^j}$. Note that

$$\sum_{i=1}^N w_{N,i}(x, z_1) c(z^j; y^i) = \widehat{C}_N(z|x) + \sum_{i=1}^N w_{N,i}(x, z_1)(c(z^j; y^i) - c(z; y^i)).$$

Moreover, by Assumptions 1 and 2, we have

$$\begin{aligned}
\mathbb{E}\left[c(z^j; Y) \middle| X = x, Z_1 = z_1\right] &= \mathbb{E}\left[c(z; Y) \middle| X = x, Z_1 = z_1\right] + \mathbb{E}\left[c(z^j; Y) - c(z; Y) \middle| X = x, Z_1 = z_1\right] \\
&= \mathbb{E}\left[c(z; Y(z_1)) \middle| X = x\right] + \mathbb{E}\left[c(z^j; Y) - c(z; Y) \middle| X = x, Z_1 = z_1\right] \\
&= C(z|x) + \mathbb{E}\left[c(z^j; Y) - c(z; Y) \middle| X = x, Z_1 = z_1\right].
\end{aligned}$$

Therefore, since $||x|| + ||z_1|| \le c_N$,

$$\begin{aligned}
\left|\widehat{C}_N(z|x) - C(z|x)\right| &= \left|\sum_{i=1}^N w_{N,i}(x, z_1) c(z; y^i) - \mathbb{E}\left[c(z; Y) \middle| X = x, Z_1 = z_1\right]\right| \\
&\le \sum_{i=1}^N w_{N,i}(x, z_1) \left|c(z; y^i) - c(z^j; y^i)\right| + \mathbb{E}\left[\left|c(z; y^i) - c(z^j; y^i)\right| \middle| X = x, Z_1 = z_1\right] \\
&\quad + \left|\sum_{i=1}^N w_{N,i}(x, z_1) c(z^j; y^i) - \mathbb{E}\left[c(z^j; Y) \middle| X = x, Z_1 = z_1\right]\right| \\
&\le \epsilon/4 + \epsilon/4 + \epsilon/2 = \epsilon.
\end{aligned}$$

Since z and ϵ were arbitrary and N_1 did not depend on z, we have

$$\tau_N = \sup_{z \in \mathcal{Z}} \left|\widehat{C}_N(z|x) - C(z|x)\right| \to 0.$$

Therefore,

$$\delta_N = \left|\inf_{z \in \mathcal{Z}} \widehat{C}_N(z|x) - \inf_{z \in \mathcal{Z}} C(z|x)\right| \le \sup_{z \in \mathcal{Z}} \left|\widehat{C}_N(z|x) - C(z|x)\right| = \tau_N \to 0.$$

Next, let $\hat{z}_N, \epsilon_N$ be such that $\epsilon_N \to 0$ and $\widehat{C}_N(\hat{z}_N|x) - \inf_{z \in \mathcal{Z}} \widehat{C}_N(z|x) \le \epsilon_N$. Then,

$$\begin{aligned}
0 \le C(\hat{z}_N|x) - \inf_{z \in \mathcal{Z}} C(z|x) &\le \left|\widehat{C}_N(\hat{z}_N|x) - C(\hat{z}_N|x)\right| + \left|\widehat{C}_N(\hat{z}_N|x) - \inf_{z \in \mathcal{Z}} \widehat{C}_N(\hat{z}_N|x)\right| \\
&\quad + \left|\inf_{z \in \mathcal{Z}} \widehat{C}_N(\hat{z}_N|x) - \inf_{z \in \mathcal{Z}} C(z|x)\right| \le \tau_N + \epsilon_N + \delta_N \to 0,
\end{aligned}$$

which completes the proof. □

Proof of Theorem 11 By assumption of Y and V sharing no atoms, $\delta \overset{a.s.}{=} \tilde{\delta} = \mathbb{I}[Y \le V]$ is observable so let us replace δ^i by $\tilde{\delta}^i$ in (24). Let

$$\begin{aligned}
F(y|x) &= \mathbb{E}\left[\mathbb{I}[Y > y] \,\middle|\, X = x\right] \\
\hat{F}_N(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y\right] w_{N,i}^{\text{Kaplan-Meier}}(x), \\
H_1(y|x) &= \mathbb{E}\left[\mathbb{I}\left[U > y, \tilde{\delta} = 1\right] \,\middle|\, X = x\right] \\
\hat{H}_{1,N}(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y,\, \tilde{\delta}^i = 1\right] w_{N,i}(x), \\
H_2(y|x) &= \mathbb{E}\left[\mathbb{I}[U > y] \,\middle|\, X = x\right] \\
\hat{H}_{2,N}(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y\right] w_{N,i}(x).
\end{aligned}$$

By assumption on conditional supports of Y and V, $\sup\{y : F(y:x) > 0\} \le \sup\{y : H_2(y:x) > 0\}$. By the same arguments as in Theorem 5, 6, 7, or 8, we have that, for all y, $\hat{H}_{1,N}(y|x) \to H_1(y|x)$, $\hat{H}_{2,N}(y|x) \to H_2(y|x)$ a.s. for μ_X-a.e.x. By assumption of conditional independence and by the main result of Beran (1981), we have that, for all y, $F_N(y|x) \to F(y|x)$ a.s. for μ_X-a.e.x. Since $\mathcal{Y}$ is a separable space we can bring the "for all y" inside the statement, i.e., we have weak convergence (c.f. Theorem 11.4.1 of Dudley (2002)): $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ a.s. for μ_X-a.e.x where $\hat{\mu}_{Y|x,N}$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Since costs are bounded, the portmanteau lemma (see Theorem 2.1 of Billingsley (1999)) gives that for each $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to \mathbb{E}[c(z;Y)|X = x]$ where $\widehat{C}_N(z|x)$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Applying Lemma EC.7 we obtain that assumptions for Lemma EC.6 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

EC.3. Extensions of Out-of-Sample Guarantees to Mixing Processes and Proofs

We can also extend the results of Section EC.1.2 to mixing processes. Combining and restating the main results of Bartlett and Mendelson (2003) (for iid) and Mohri and Rostamizadeh (2008) (for mixing), we can restate Theorem EC.3 as follows

THEOREM EC.10. *Consider a class $\mathcal{G}$ of functions $\mathcal{U} \to \mathbb{R}$ that are bounded: $|g(u)| \le \overline{g}$ $\forall g \in \mathcal{G}$, $u \in \mathcal{U}$. Consider a sample $S_n = (u^1, \ldots, u^N)$ of some random variable $T \in \mathcal{T}$. Fix $\delta > 0$. If S_N is generated by IID sampling, let $\delta' = \delta'' = \delta$ and $\nu = N$. If S_N comes from a β-mixing process, fix some t, ν such that $2t\nu = N$, let $\delta' = \delta/2 - (\nu-1)\beta(t)$ and $\delta'' = \delta/2 - 2(\nu-1)\beta(t)$. Then (only for $\delta' > 0$ or $\delta'' > 0$ where they appear), we have that with probability $1-\delta$,*

$$\mathbb{E}[g(T)] \le \frac{1}{N}\sum_{i=1}^N g(u^i) + \overline{g}\sqrt{\log(1/\delta')/2\nu} + \mathfrak{R}_\nu(\mathcal{G}) \qquad \forall g \in \mathcal{G}, \tag{EC.8}$$

and that, again, with probability $1-\delta$,

$$\mathbb{E}\left[g(T)\right] \leq \frac{1}{N}\sum_{i=1}^{N} g(u^i) + 3\overline{g}\sqrt{\log(2/\delta')/2\nu} + \widehat{\mathfrak{R}}_\nu(\mathcal{G}) \qquad \forall g \in \mathcal{G}. \tag{EC.9}$$

Replacing this result in the proof of Theorem EC.2 extends it to the case of data generated by a mixing process.

EC.4. Proofs of Tractability Results

Proof of Theorem 2 Let $I = \{i : w_{N,i}(x) > 0\}$, $w = (w_{N,i}(x))_{i\in I}$. Rewrite (3) as $\min w^T\theta$ over $(z,\theta) \in \mathbb{R}^{d\times n_0}$ subject to $z \in \mathcal{Z}$ and $\theta_i \geq c(z;y^i)$ $\forall i \in I$. Weak optimization of a linear objective over a closed convex body is reducible to weak separation via the ellipsoid algorithm (see Grotschel et al. (1993)). A weak separation oracle for $\mathcal{Z}$ is assumed given. To separate over the i^{th} cost constraint at fixed z', θ_i' call the evaluation oracle to check violation and if violated call the subgradient oracle to get $s \in \partial_z c(z';y^i)$ with $||s||_\infty \leq 1$ and produce the cut $\theta_i \geq c(z';y^i) + s^T(z-z')$. □

Proof of Theorem 3 Solve (21) for each of $z_{11},\ldots,z_{1b}$ and take the minimum. In each case, we have a problem that resembles (3) and we may use an argument similar to Theorem 2 to prove its tractability. □

Proof of Theorem 4 Let $R = \inf_{z_1'\in\mathcal{Z}_1}\sup_{z_1\in\mathcal{Z}_1}||z_1 - z_1'||$. Then no more than $b = (3RL/\epsilon)^{d_{z_1}}$ balls of radius ϵ/L are needed in order to cover $\mathcal{Z}_1$. Let their centers be denoted $z_{11},\ldots,z_{1b}$ and apply Theorem 3. □

Proof of Theorem EC.1 In the case of (EC.1), $z(x^i) = Wx^i$. By computing the norm of W we have a trivial weak membership algorithm for the norm constraint and hence by Theorems 4.3.2 and 4.4.4 of Grotschel et al. (1993) we have a weak separation algorithm. By adding affine constraints $\zeta_{ij} = z_j(x^i)$, all that is left is to separate over are constraints of the form $\theta_i \geq c(\zeta_i;y^i)$, which can be done as in the proof of Theorem 2. □

EC.5. Proof of Theorem 1

Proof Under Assumptions 1 and 2, the objective of problem (19) can be rewritten as

$$\begin{aligned}
\mathbb{E}\left[c(z;Y(z))\middle|X=x\right] &= \mathbb{E}\left[c(z;Y(z_1))\middle|X=x\right] && \text{(By Assumption 1)}\\
&= \mathbb{E}\left[c(z;Y(z_1))\middle|X=x, Z_1=z_1\right] && \text{(By Assumption 2)}\\
&= \mathbb{E}\left[c(z;Y(Z_1))\middle|X=x, Z_1=z_1\right] && \text{(By conditioning)}\\
&= \mathbb{E}\left[c(z;Y)\middle|X=x, Z_1=z_1\right] && \text{(By definition of } Y\text{)}
\end{aligned}$$

which is the objective of problem (20). □

EC.6. Omitted Details from Sections 1.1 and 3.2

EC.6.1. Shipment Planning Example

In our shipment planning example, we consider stocking $d_z = 4$ warehouses to serve $d_y = 12$ locations. We take locations spaced evenly on the 2-dimensional unit circle and warehouses spaced evenly on the circle of radius 0.85. The resulting network and its associated distance matrix are shown in Figure EC.1. We suppose shipping costs from warehouse i to location j are $c_{ij} = \$10 D_{ij}$ and that production costs are \$5 per unit when done in advance and \$100 per unit when done last minute.

We consider observing $d_x = 3$ demand-predictive features X that, instead of iid, evolve as a 3-dimensioanl ARMA(2,2) process:

$$X(t) - \Phi_1 X(t-1) - \Phi_2 X(t-2) = U(t) + \Theta_1 U(t-1) + \Theta_2 U(t-2)$$

where $U \sim \mathcal{N}(0, \Sigma_U)$ are innovations and

$$\Phi_1 = \begin{pmatrix} 0.5 & -0.9 & 0 \\ 1.1 & -0.7 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0. & -0.5 & 0 \\ -0.5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

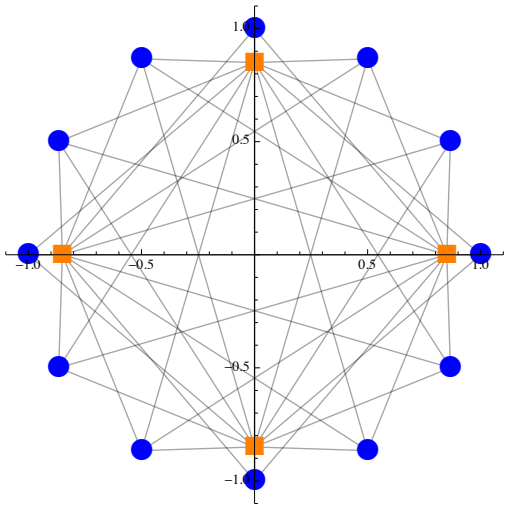
$$\Theta_1 = \begin{pmatrix} 0.4 & 0.8 & 0 \\ -1.1 & -0.3 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Theta_2 = \begin{pmatrix} 0 & -0.8 & 0 \\ -1.1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Sigma_U = \begin{pmatrix} 1. & 0.5 & 0 \\ 0.5 & 1.2 & 0.5 \\ 0 & 0.5 & 0.8 \end{pmatrix}.$$

We suppose that demands are generated according to a factor model

$$Y_i = \max\{0,\, A_i^T (X + \delta_i/4) + \left(B_i^T X\right) \epsilon_i\}$$

where A_i is the mean-dependence of the i^{th} demand on these factors with some idiosyncratic noise, B_i the variance-dependence, and ϵ_i and δ_i are independent standard Gaussian idiosyncratic contributions. For A and B we use

$$A = 2.5 \times \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \quad B = 7.5 \times \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$



(a) The network of warehouses (orange) and locations (blue).

$$D^T = \begin{pmatrix} 0.15 & 1.3124 & 1.85 & 1.3124 \\ 0.50026 & 0.93408 & 1.7874 & 1.6039 \\ 0.93408 & 0.50026 & 1.6039 & 1.7874 \\ 1.3124 & 0.15 & 1.3124 & 1.85 \\ 1.6039 & 0.50026 & 0.93408 & 1.7874 \\ 1.7874 & 0.93408 & 0.50026 & 1.6039 \\ 1.85 & 1.3124 & 0.15 & 1.3124 \\ 1.7874 & 1.6039 & 0.50026 & 0.93408 \\ 1.6039 & 1.7874 & 0.93408 & 0.50026 \\ 1.3124 & 1.85 & 1.3124 & 0.15 \\ 0.93408 & 1.7874 & 1.6039 & 0.50026 \\ 0.50026 & 1.6039 & 1.7874 & 0.93408 \end{pmatrix}$$

(b) The distance matrix.

Figure EC.1 Network data for shipment planning example.

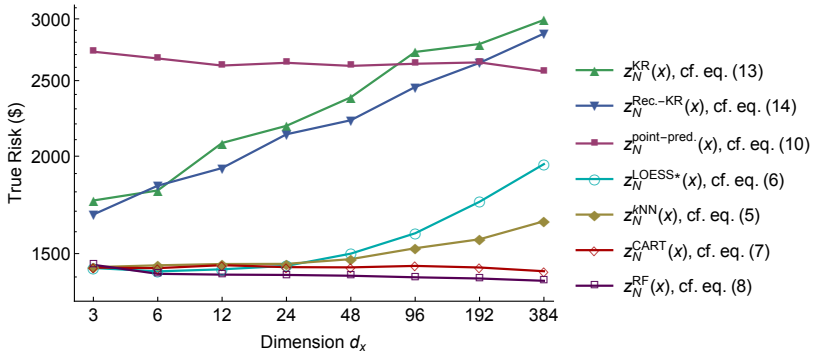


Figure EC.2 Results of the experiment in Section EC.6.2.

EC.6.2. Effect of Additional Dimensions with Diminishing Predictiveness

In Section 1.1, we presented an experiment to demonstrate the effect of increasing dimension on various predictive prescriptions. In the experiment, added dimensions were noise that was a priori not distinguishable from the three main features. We can consider an alternative set up to the experiment to investigate the effect of increasing dimension when added dimensions provide additional, marginal increase in the predictiveness of Y. Toward that end, for each $L \in \mathbb{N}$, we consider $3L$ auxiliary variables consisting of L copies of the original variables in the example where the ℓ^{th} copy is transformed by adding standard normal noise to the 3 variables multiplied by $(1/2)^\ell$. Thus, each additional copy can be used to better pin down the original variables but is more noisy than the last copy. As in Section 1.1 we fix $N = 2^{14}$. For each $L = 1, 2, 4, \ldots, 128$, we rerun the example with these alternative variables and plot the results in Figure EC.2.

There are a few things to note about the results. First, the performance of the even best predictive prescriptions for $d_x = 3$ suffers due to the added noise of the first copy (compare to $d_x = 3$ in Figure 1b). Second, the performance of both the point-prediction-driven decision (using RF) and the predictive prescriptions based on CART and RF improves slowly as dimension grows and these methods use the marginal amounts of additional predictive power in the data. This shows a difference to the setting in Section 1.1, where these performed the same or very slightly worse as dimension grew. Of course, these predictive prescriptions we develop do much better than the point-prediction-driven decision, which only beats the worst ones (KR and Rec.-KR) in very high dimensions. Third, the performance of the predictive prescriptions based on KR, Rec.-KR, LOESS*, and kNN all have performance that deteriorates with dimension due to the curse of dimensionality and these methods' inability to learn which features are more important than others, but the deterioration is milder than in Section 1.1 and depends on the balance between the growth of dimension and the additional predictiveness offered.

EC.6.3. Shipment Planning with Pricing Example

In our shipment planning with pricing example from Section 3.2, we used the same parameters as in the above shipment planning example, except that we set

$$Y_i(z_1) = \frac{100}{1+e^{\frac{z_1-50}{100}}} \max\{0,\, A_i^T\left(X+\delta_i/4\right) + \left(B_i^T X\right)\epsilon_i\},$$

and we simulate the historical price data log-normally as

$$\log(Z_1) \sim \mathcal{N}(||X||_1/5,\, \left|400 + e^T X\right|),$$

where e is the vector of all ones.

EC.6.4. Varying the Determination of Y

To vary the determination of Y as in Section 5 we do as follows. We let $X' \in \mathbb{R}^{d_x}$ be normally distributed with 0 mean and covariance matrix equal to the covariance of X, Σ_X. We then introduce a new parameter $\kappa \in [0,2]$, let $\kappa' = \max\{0, 1-\kappa\}$ and $\kappa'' = \min\{1, 2-\kappa\}$ and redefine

$$Y_i = \max\{0,\, A_i^T\left(\sqrt{\kappa}X + \sqrt{\kappa'}X' + \sqrt{\kappa''}\delta_i/4\right) + \sqrt{\kappa''}\sqrt{\kappa}\left(B_i^T X\right)\epsilon_i + \sqrt{\kappa''}\sqrt{\kappa'}(V\Sigma_X V^T)_i^{T/2}\epsilon\}.$$

The original example corresponds to $\kappa = 1$, whereas $\kappa = 0$ corresponds to independence between X and Y and $\kappa = 2$ corresponds to Y being measurable with respect to (a function of) X.

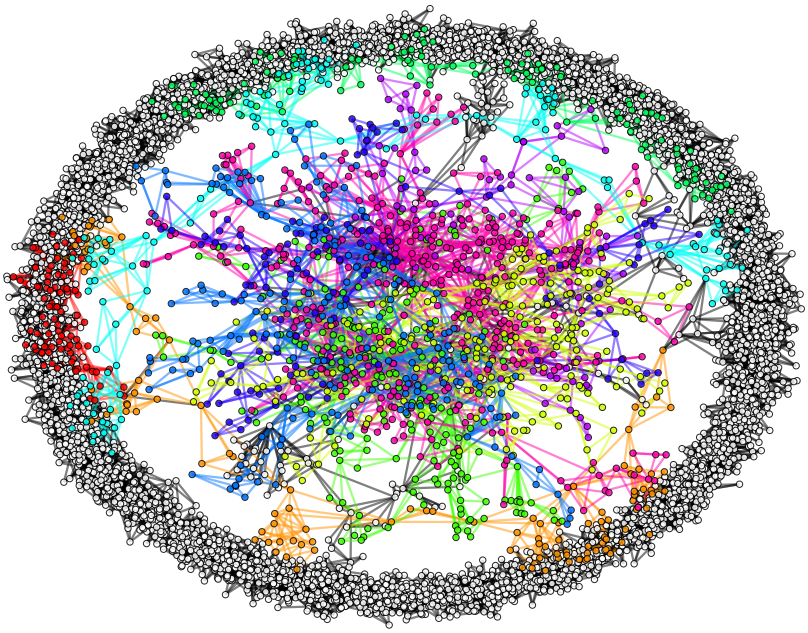


Figure EC.3 The graph of actors, connected via common movies where both are first-billed. Colored nodes correspond to the 10 largest communities of actors. Colored edges correspond to intra-community edges.

EC.7. Constructing Auxiliary Data Features from IMDb and RT data

For some information harvested from IMDb and RT, the corresponding numeric feature is straightforward (e.g. number of awards). For other pieces of information, some distillation is necessary. For genre, we create an indicator vector. For MPAA rating, we create a single ordinal (from 1 for G to 5 for NC-17). For plot, we measure the cosine-similarity between plots,

$$\text{similarity}(P_1, P_2) = \frac{p_1^T p_2}{||p_1||\,||p_2||},$$

where p_{ki} denotes the number of times word i appears in plot text P_k and i indexes the collection of unique words appearing in plots P_1, P_2 ignoring certain generic words like "the". and use this as a distance measure to hierarchically cluster the plots using Ward's method (cf. Ward (1963)). This captures common themes in titles. We construct 12 clusters based solely on historical data and, for new data, include a feature vector of median cosine similarity to each of the clusters. For actors, we create a graph with titles as nodes and with edges between titles that share actors, weighted by the number of actors shared. We use the method of Blondel et al. (2008) to find communities of titles and create an actor-counter vector for memberships in the 10 largest communities (see Figure EC.3). This approach is motivated by the existence of such actor groups as the "Rat Pack" (Humphery Bogart and friends), "Brat Pack" (Molly Ringwald and friends), and "Frat Pack" (Owen Wilson and friends) that often co-star in titles with a similar theme, style, and target audience.

EC References

Bartlett, Peter, Shahar Mendelson. 2003. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482.

Beran, Rudolf. 1981. Nonparametric regression with randomly censored survival data. Tech. rep.

Bertsekas, Dimitri, Angelia Nedić, Asuman Ozdaglar. 2003. *Convex analysis and optimization*. Athena Scientific, Belmont.

Bertsekas, Dimitri P. 1999. *Nonlinear programming*. Athena Scientific, Belmont.

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory E.* **2008**(10) P10008.

Bradley, Richard. 1986. Basic properties of strong mixing conditions. *Dependence in Probability and Statistics*. Birkhausser, 165–192.

Bradley, Richard. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2**(107-44) 37.

Carrasco, Marine, Xiaohong Chen. 2002. Mixing and moment properties of various garch and stochastic volatility models. *Economet. Theor.* **18**(1) 17–39.

Doukhan, Paul. 1994. *Mixing: Properties and Examples*. Springer.

Dudley, Richard M. 2002. *Real analysis and probability*, vol. 74. Cambridge University, Cambridge.

Geer, Sara A. 2000. *Empirical Processes in M-estimation*. Cambridge University.

Grotschel, M, L Lovasz, A Schrijver. 1993. *Geometric algorithms and combinatorial optimization*. Springer, New York.

Hansen, Bruce E. 2008. Uniform convergence rates for kernel estimation with dependent data. *Economet. Theor.* **24**(03) 726–748.

Kakade, Sham, Karthik Sridharan, Ambuj Tewari. 2008. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *NIPS*. 793–800.

Koenker, Roger. 2005. *Quantile regression*. 38, Cambridge University.

Ledoux, Michel, Michel Talagrand. 1991. *Probability in Banach Spaces: isoperimetry and processes*. Springer.

Mcdonald, Daniel, Cosma Shalizi, Mark Schervish. 2011. Estimating beta-mixing coefficients. *AISTATS*. 516–524.

Mohri, Mehryar, Afshin Rostamizadeh. 2008. Rademacher complexity bounds for non-iid processes. *NIPS*. 1097–1104.

Mokkadem, Abdelkader. 1988. Mixing properties of arma processes. *Stoch. Proc. Appl.* **29**(2) 309–315.

Rudin, Cynthia, Gah-Yi Vahn. 2015. The big data newsvendor: Practical insights from machine learning.

Walk, Harro. 2010. Strong laws of large numbers and nonparametric estimation. *Recent Developments in Applied Probability and Statistics*. Springer, 183–214.

Ward, Joe. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**(301) 236–244.