CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 13: Representation learning and transfer

October 15, 2026

Lecturer: Csaba Szepesvári *Note prepared by: Csaba Szepesvári*

Reading guide. Neural networks, including language models, are often said to be powerful because they learn representations. This is usually contrasted with "linear" methods where a feature map needs to be hand-picked for the method to work well.

We first define representation-learning algorithms in the context of learning a map and connect the definition to neural networks. Representation learning has a clear formulation in the multitask setting: learn one feature map from several tasks, with a separate predictor for each task. We ask whether sharing reduces the samples needed to predict on those tasks. Transfer learning asks a further question: do the learned features help a new task? We give a finite-class analysis of each problem, then an efficient transfer algorithm for a linear model. The statistical and computational questions are separate: how much data suffices, and how can the learner compute useful features?

The main results are the joint-ERM bounds in Theorems 13.2.2 and 13.2.4, the separation from independent learning in Proposition 13.2.3, and the transfer guarantees in Theorem 13.3.2 and Corollary 13.3.6. The proofs use the finite-class argument from Lecture 2 and concentration from Lecture 6. The examples also connect to parameter sharing in Lecture 4 and coverage in Lecture 7.

13.1 Representation learning

Let $\mathcal{X}$ be an input space, $\mathcal{Y}$ a label space, and $\mathcal{Z}$ a feature space. We consider learning a predictor $\mathcal{X} \to \mathcal{Y}$ from n labeled examples.

Definition 13.1.1 (Representation-learning algorithm)**.** Let $\mathcal{H}$ be a class of maps $\mathcal{X} \to \mathcal{Z}$ and $\mathcal{G}$ a class of maps $\mathcal{Z} \to \mathcal{Y}$. A *representation-learning algorithm* for sample size n with representation class $\mathcal{H}$ and head class $\mathcal{G}$ is a map

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{H} \times \mathcal{G}.$$

Writing $\mathcal{A}(S) = (h_S, g_S)$, the learner returns a *representation* (map) h_S and a *head* g_S. The resulting predictor is $\widehat{f}_S = g_S \circ h_S$. For example, h_S can compute a neural network's hidden features and g_S can classify using those features. As can be seen from the definition, the main distinctive feature of a representation learner is the signature of the underlying map. Namely, that it returns a map that is explicitly decomposed into two maps. This definition is meant to capture the essence of these methods and as such is a bit oversimplified. Every supervised learning algorithm is a trivial representation learner: Just choose $\mathcal{Z} = \mathcal{X}$, $\mathcal{H}$ consist of the $\mathcal{X} \to \mathcal{X}$ identity map. Of course, we are more interested in nontrivial representation learners, for which for at least some sample $S \in (\mathcal{X} \times \mathcal{Y})^n$, h_S is not the identity map (if $\mathcal{Z} = \mathcal{X}$).

Our primary interest in representation learning is to gain insight into why learning with neural networks is special as compared to simpler alternatives. To explain this connection recall that a feedforward neural network is given by a directed (computation) graph, with nodes performing

simple computation over the information sent to them along the edges ending at them, where learning changes the actual computation performed. Now, take any minimum-size cut that separates the inputs (sources of the graph) from the outputs (sinks of the graph): This is a set of nodes such that removing this set, no input is connected to no output, while no subset of this set of nodes has the same property. If the nodes form a layered graph (with respect to distance to inputs), as it happens for example, in multi-layered networks, any node set in a layer will be such cut. The map determined by the part of the computation graph up to the cut defines $h : x \mapsto z$; the remaining computation defines $g : z \mapsto \widehat{y}$, while the whole network computes $g \circ h$.

It follows that neural networks are in a way "super representation learners": They are considering many compositional maps simultaneously (corresponding to all possible cuts) and I suspect this is one of their strengths. However, for the sake of this lecture we stick to a simpler view that ignores this multi-representation learning aspect, focusing on just a one-way decomposition of the target function.

13.2 Multitask learning

The usefulness of representation learning is easiest to argue for when the learner needs to solve multiple prediction problems, that share a common representation. In this section we study this problem in *multitask learning*, where a learner needs to find predictors for several tasks and is evaluated on those tasks. We extend the learner's output to one representation and T heads. The shared-map, separate-head structure makes it easy to articulate the potential benefits of learning a shared representation.

13.2.1 The learning problem

The learner receives labeled data from each task of the $t \in [T]$ tasks and returns a predictor for each task. For the first statistical comparison, take $\mathcal{H}$ and $\mathcal{G}$ to be finite and nonempty. The finite-class model isolates the accounting created by a shared representation: the representation complexity is paid once across tasks, while the head complexity is paid once per task. More general representation-class bounds replace logarithmic cardinalities by other complexity measures and retain this separation. Such class-level bounds characterize the statistical behavior of the class. Explaining which representation neural-network training selects from a large architecture class additionally requires properties of the training procedure and the learned representation. The same selection problem is central to neural-network generalization theory beyond multitask and representation learning. In the Gaussian linear model studied later, we analyze an explicit representation-learning strategy and identify what an algorithm-level guarantee requires.

Let $P_1, \dots, P_T$ on $\mathcal{X} \times \mathcal{Y}$ be the data-generating laws of the individual T tasks. For each task $t \in [T]$, the sample is

$$S_t = ((X_{t1}, Y_{t1}), \dots, (X_{tn}, Y_{tn})) \sim P_t^{\otimes n}.$$

All nT pairs are independent. The task identity of each example is given.[1] The laws may differ, including their input marginals. No statistical model over the tasks is assumed.

Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a loss. Write

$$R_t(f) = \int \ell(f(x), y)\, P_t(dx, dy), \qquad R_{S_t}(f) = \frac{1}{n} \sum_{j=1}^{n} \ell(f(X_{tj}), Y_{tj}). \tag{13.1}$$

A sample in the subscript marks an empirical risk: R_t is computed from the law P_t, while R_{S_t} is computed from the sample S_t. Write $S = (S_1, \dots, S_T)$ for the entire source sample. For an output

$(h, \boldsymbol{g}) \in \mathcal{H} \times \mathcal{G}^T$, with $\boldsymbol{g} = (g_1, \dots, g_T)$, the task-$t$ predictor is $g_t \circ h$. We evaluate the output by its average population risk and use the corresponding empirical risk for training:

$$\overline{R}(h, \boldsymbol{g}) = \frac{1}{T} \sum_{t=1}^{T} R_t(g_t \circ h), \qquad \overline{R}_S(h, \boldsymbol{g}) = \frac{1}{T} \sum_{t=1}^{T} R_{S_t}(g_t \circ h). \tag{13.2}$$

A bar marks an average over the T source tasks, and a hat marks a quantity returned by a learner, such as the pair $(\widehat{h}, \widehat{\boldsymbol{g}})$ defined next. Joint empirical risk minimization chooses one representation and all the heads:

$$(\widehat{h}, \widehat{\boldsymbol{g}}) \in \operatorname*{argmin}_{h \in \mathcal{H},\ \boldsymbol{g} \in \mathcal{G}^T} \overline{R}_S(h, \boldsymbol{g}). \tag{13.3}$$

Independent learning instead chooses a separate (h_t, g_t) for every task.

Whether joint learning offers a statistical advantage over independent learning depends on the head class $\mathcal{G}$. An unrestricted head could undo the statistical restriction imposed by h. The intended regime has a rich representation class and a relatively simple head class: $\ln |\mathcal{H}|$ is large compared with $\ln |\mathcal{G}|$.

13.2.2 The statistical benefit of sharing

We first compare joint and separate learning when one representation permits perfect prediction on all tasks. We then allow an approximation cost for requiring a common representation.

13.2.2.1 Realizable bounded losses

Suppose one representation $h^\star \in \mathcal{H}$ and heads $g_1^\star, \dots, g_T^\star \in \mathcal{G}$ achieve zero risk on every source task. Because the loss is nonnegative, every source loss of this predictor tuple equals zero almost surely. Joint ERM therefore has zero empirical loss almost surely. The consistency argument from Theorem 2.7.4 extends as follows to the bounded loss from Section 13.2.1 and to independent observations whose laws may differ.

Let $\mathcal{F}$ be a finite nonempty class of predictors from $\mathcal{U}$ to $\mathcal{Y}$. Let $(U_1, V_1), \dots, (U_m, V_m)$ be independent, not necessarily identically distributed, and let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. For $f \in \mathcal{F}$, define

$$R_i(f) = \mathbb{E}[\ell(f(U_i), V_i)], \qquad \overline{R}_m(f) = \frac{1}{m} \sum_{i=1}^{m} R_i(f).$$

Suppose some $f^\star \in \mathcal{F}$ satisfies $R_i(f^\star) = 0$ for every $i \in [m]$. On the same experiment, let $\widehat{f}$ be an $\mathcal{F}$-valued random predictor with

$$\ell(\widehat{f}(U_i), V_i) = 0 \qquad \text{for every } i \in [m]$$

almost surely. Such a predictor exists because $R_i(f^\star) = 0$ and the loss is nonnegative.

Proposition 13.2.1 (Finite-class zero-loss bound)**.** *For every $\delta \in (0, 1)$,*

$$\mathbb{P}\left\{ \overline{R}_m(\widehat{f}) \le \frac{\ln |\mathcal{F}| + \ln(1/\delta)}{m} \right\} \ge 1 - \delta.$$

Moreover,

$$\mathbb{E}[\overline{R}_m(\widehat{f})] \le \frac{\ln |\mathcal{F}| + 1}{m}.$$

Exercise 1 asks you to prove the proposition. The key inequality is

$$\mathbb{P}\big(\ell(f(U_i), V_i) = 0\big) \le 1 - R_i(f),$$

which follows from $0 \le \ell \le 1$. For 0–1 loss, the inequality is an equality and the proposition reduces to the usual consistency bound.

To apply the proposition, augment each input with its task identity and consider the class

$$\mathcal{F}_{\text{joint}} = \{(t, x) \mapsto g_t(h(x)) : h \in \mathcal{H},\ (g_1, \dots, g_T) \in \mathcal{G}^T\}.$$

The class has cardinality at most $|\mathcal{H}||\mathcal{G}|^T$. For the nT augmented observations $((t, X_{tj}), Y_{tj})$, the quantity $\overline{R}_{nT}$ in Proposition 13.2.1 equals $\overline{R}(h, \boldsymbol{g})$. The subscripts src and trg used below abbreviate source and target. For $\delta \in (0, 1)$, put

$$a_{\text{src}}(\delta) = \frac{\ln|\mathcal{H}| + T\ln|\mathcal{G}| + \ln(1/\delta)}{nT}.$$

Theorem 13.2.2 (Joint zero-loss bound)**.** *Under the zero-risk source assumption above, let $(\widehat{h}, \widehat{\boldsymbol{g}})$ be any predictor tuple with zero loss on every source example. For $0 < \delta < 1$,*

$$\mathbb{P}\left\{\overline{R}(\widehat{h}, \widehat{\boldsymbol{g}}) \le a_{\text{src}}(\delta)\right\} \ge 1 - \delta. \tag{13.4}$$

The expected average risk also satisfies

$$\mathbb{E}[\overline{R}(\widehat{h}, \widehat{\boldsymbol{g}})] \le \frac{\ln|\mathcal{H}| + T\ln|\mathcal{G}| + 1}{nT}. \tag{13.5}$$

Proof. Apply Proposition 13.2.1 with $m = nT$ and $|\mathcal{F}_{\text{joint}}| \le |\mathcal{H}||\mathcal{G}|^T$. □

The bound separates the two statistical costs:

$$a_{\text{src}}(\delta) = \underbrace{\frac{\ln|\mathcal{H}|}{nT}}_{\text{one shared representation}} + \underbrace{\frac{\ln|\mathcal{G}|}{n}}_{\text{one head per task}} + \frac{\ln(1/\delta)}{nT}.$$

For the comparison below, write

$$\overline{R}_{\text{shared}} := \overline{R}(\widehat{h}, \widehat{\boldsymbol{g}})$$

for the risk of the shared learner's output. Independent learning returns a separate pair $(\widehat{h}_t^{\text{ind}}, \widehat{g}_t^{\text{ind}}) \in \mathcal{H} \times \mathcal{G}$ for each task t; define its average risk by

$$\overline{R}_{\text{independent}} := \frac{1}{T}\sum_{t=1}^{T} R_t(\widehat{g}_t^{\text{ind}} \circ \widehat{h}_t^{\text{ind}}).$$

The class of independent predictor tuples has size at most $(|\mathcal{H}||\mathcal{G}|)^T$. The same finite-class result gives the sufficient bound

$$\mathbb{P}\left\{\overline{R}_{\text{independent}} \le \frac{\ln|\mathcal{H}| + \ln|\mathcal{G}|}{n} + \frac{\ln(1/\delta)}{nT}\right\} \ge 1 - \delta. \tag{13.6}$$

Thus the representation term in the shared upper bound is smaller by a factor of T. Comparing upper bounds is a useful first comparison. The next example also gives a lower bound for independent learning.

Lecture 4 analyzed pooling across positions when parameters are shared. Here the pooled object is a feature map rather than a single probability. In both cases the statistical gain depends on a restriction that is correct for the tasks being pooled.

13.2.2.2 An example where sharing provably helps

Return to Lecture 2's boxes. Let $\mathcal{X} = [d]$ with $d \geq 2$. An unknown map $h^\star : [d] \to \{0, 1\}$ assigns a bit to every box. Task t has an unknown bit b_t and label

$$Y = h^\star(X) \oplus b_t, \qquad X \text{ uniform on } [d], \tag{13.7}$$

where $\oplus$ denotes addition modulo two. All training inputs are independent. Choose all 2^d Boolean maps as $\mathcal{H}$ and the two heads $g_b(z) = z \oplus b$ as $\mathcal{G}$.

Proposition 13.2.3 (Shared boxes versus independent learning)**.** *Joint ERM in this model has*

$$\mathbb{E}[\overline{R}_{\text{shared}}] \leq \frac{(d+T)\ln 2 + 1}{nT}. \tag{13.8}$$

For any collection of learners whose task-t predictor uses only S_t and independent randomization, there is a fixed choice of $(h^\star, b_1, \ldots, b_T)$ for which

$$\mathbb{E}[\overline{R}_{\text{independent}}] \geq \frac{1}{2}\left(1 - \frac{1}{d}\right)^n. \tag{13.9}$$

Proof. The upper bound is Eq. (13.5). For the lower bound, set every $b_t = 0$ and temporarily draw the d values of $h^\star$ independently as fair bits, independently of all inputs and learner randomization. A fresh uniform test box is absent from S_t with probability $(1 - 1/d)^n$. Conditional on that event and the task-t data, its bit remains fair. Every task-t prediction therefore has conditional error $1/2$ on the event. Average over tasks and then over the random map. At least one fixed map has average expected error at least the displayed lower bound. □

For a fixed error tolerance $\varepsilon > 0$, the shared upper bound shows that

$$n \geq \left\lceil \frac{(d+T)\ln 2 + 1}{\varepsilon T} \right\rceil$$

examples per task suffice to make $\mathbb{E}[\overline{R}_{\text{shared}}] \leq \varepsilon$. Thus $O(1 + d/T)$ examples per task suffice for any fixed tolerance. In contrast, independent learning needs $n = \Omega(d)$ to guarantee a fixed small tolerance uniformly over this family. For example, Bernoulli's inequality gives an independent-learning error of at least $1/4$ whenever $n \leq d/2$. When T is of order d, the shared upper bound can be made smaller than $1/4$ with a constant n. These are statistical statements; enumerating all representations need not be computationally efficient.

13.2.2.3 Misspecification and joint ERM

We now allow arbitrary task laws and any loss in $[0, 1]$. Define

$$R_t^\star = \min_{h \in \mathcal{H},\ g \in \mathcal{G}} R_t(g \circ h), \qquad A_t(h) = \min_{g \in \mathcal{G}} R_t(g \circ h) - R_t^\star. \tag{13.10}$$

The quantity $A_t(h) \geq 0$ is the extra risk imposed by using representation h, even with its best head. The benchmark $R_t^\star$ allows task t its own representation. It need not equal Bayes risk: the composite class itself may be misspecified.

Write

$$\overline{A}(h) = \frac{1}{T}\sum_{t=1}^{T} A_t(h), \qquad \Delta_{\text{share}} = \min_{h \in \mathcal{H}} \overline{A}(h). \tag{13.11}$$

The number Δ_{share} is the approximation cost of requiring one representation. Because the heads can be chosen separately,

$$\min_{h,\boldsymbol{g}} \overline{R}(h, \boldsymbol{g}) = \frac{1}{T} \sum_{t=1}^{T} R_t^\star + \Delta_{\text{share}}. \tag{13.12}$$

Exact sharing means $\Delta_{\text{share}} = 0$, not necessarily zero label noise.

Allow the source learner an optimization error $\eta_{\text{src}} \geq 0$:

$$\overline{R}_S(\widehat{h}, \widehat{\boldsymbol{g}}) \leq \min_{h,\boldsymbol{g}} \overline{R}_S(h, \boldsymbol{g}) + \eta_{\text{src}}. \tag{13.13}$$

Theorem 13.2.4 (Agnostic joint ERM)**.** *Under Section 13.2.1 and Eq. (13.13), define*

$$b_{\text{src}}(\delta) = \sqrt{\frac{2\{\ln|\mathcal{H}| + T \ln|\mathcal{G}| + \ln(2/\delta)\}}{nT}}.$$

With probability at least $1 - \delta$*,*

$$\overline{R}(\widehat{h}, \widehat{\boldsymbol{g}}) - \frac{1}{T} \sum_{t=1}^{T} R_t^\star \leq \Delta_{\text{share}} + \eta_{\text{src}} + b_{\text{src}}(\delta), \tag{13.14}$$

$$\overline{A}(\widehat{h}) \leq \Delta_{\text{share}} + \eta_{\text{src}} + b_{\text{src}}(\delta). \tag{13.15}$$

Proof. For a fixed tuple, Hoeffding's inequality applies to the nT independent, bounded losses, although their laws differ across tasks. A union bound over at most $K = |\mathcal{H}||\mathcal{G}|^T$ tuples gives

$$\sup_{h,\boldsymbol{g}} |\overline{R}_S(h, \boldsymbol{g}) - \overline{R}(h, \boldsymbol{g})| \leq e_{\text{src}} := \sqrt{\frac{\ln(2K/\delta)}{2nT}}$$

with probability at least $1 - \delta$. On this event, for a population minimizer $(h^\circ, \boldsymbol{g}^\circ)$,

$$\begin{aligned} \overline{R}(\widehat{h}, \widehat{\boldsymbol{g}}) &\leq \overline{R}_S(\widehat{h}, \widehat{\boldsymbol{g}}) + e_{\text{src}} \\ &\leq \overline{R}_S(h^\circ, \boldsymbol{g}^\circ) + \eta_{\text{src}} + e_{\text{src}} \\ &\leq \overline{R}(h^\circ, \boldsymbol{g}^\circ) + \eta_{\text{src}} + 2e_{\text{src}}. \end{aligned}$$

Use Eq. (13.12) and $2e_{\text{src}} = b_{\text{src}}(\delta)$ for the first claim. For the second, minimize each population head risk at $\widehat{h}$: $\overline{A}(\widehat{h}) \leq \overline{R}(\widehat{h}, \widehat{\boldsymbol{g}}) - T^{-1} \sum_t R_t^\star$. □

The estimation term has order $\sqrt{\ln|\mathcal{H}|/(nT) + \ln|\mathcal{G}|/n}$. Independent ERM has an analogous bound with order $\sqrt{(\ln|\mathcal{H}| + \ln|\mathcal{G}|)/n}$ and no Δ_{share}. Sharing reduces the estimation cost but can increase approximation error. When $\Delta_{\text{share}} > 0$, even the best shared predictor has greater average risk than the best separate predictors. Whether sharing hurts at a finite sample size depends on the estimation errors as well. The added approximation error is one source of *negative transfer*: sharing can increase risk relative to learning separately.

13.3 Transfer learning

In *transfer learning*, information learned from source tasks is used to learn a predictor for a target task. We study a two-stage procedure: estimate $\widehat{h}$ from the source samples $S_1, \ldots, S_T$, then freeze $\widehat{h}$ and fit a head using examples from a new target task. Learning how to fit new tasks from experience on source tasks is also called *learning-to-learn* or *meta-learning*.

Fix a target law P_0 and define R_0, $R_0^\star$, and $A_0(h)$ as in Eqs. (13.1) and (13.10). After estimating $\widehat{h}$, the learner receives $S_0 = ((X_{01}, Y_{01}), \dots, (X_{0m}, Y_{0m})) \sim P_0^{\otimes m}$, independently of all source data and source-learner randomization. The output $(\widehat{h}, \widehat{g}_0)$ predicts by $\widehat{f}_0 = \widehat{g}_0 \circ \widehat{h}$. The objective is small target excess risk $R_0(\widehat{f}_0) - R_0^\star$.

13.3.1 Finite-class analysis

We retain the finite classes and bounded loss from Section 13.2.1. Theorem 13.2.4 bounds $\overline{A}(\widehat{h})$. A guarantee for a particular task requires the source tasks to constrain the features needed by that task. The existence of one representation that works for all tasks leaves the behavior of the learned representation undetermined.

The source tasks constrain a candidate representation through its average source representation error. Define the *transfer envelope*

$$\Psi(u) = \max\big(\{A_0(h) : h \in \mathcal{H},\ \overline{A}(h) \le u\} \cup \{0\}\big), \qquad u \ge 0. \tag{13.16}$$

The value $\Psi(u)$ is the largest target representation error among the candidate representations whose average source representation error is at most u. Thus Ψ states exactly what source performance permits on the target. The envelope is nondecreasing and, by construction,

$$A_0(h) \le \Psi(\overline{A}(h)) \qquad \text{for every } h \in \mathcal{H}. \tag{13.17}$$

Proposition 13.3.1 (Sharpness of the transfer envelope)**.** *Let $\phi : [0, \infty) \to [0, \infty)$ be nondecreasing and suppose*

$$A_0(h) \le \phi(\overline{A}(h)) \qquad \textit{for every } h \in \mathcal{H}.$$

Then $\Psi(u) \le \phi(u)$ for every $u \ge 0$. Moreover, whenever $\{h \in \mathcal{H} : \overline{A}(h) \le u\}$ is nonempty, some $h_u \in \mathcal{H}$ satisfies

$$\overline{A}(h_u) \le u, \qquad A_0(h_u) = \Psi(u).$$

Exercise 3 asks you to prove the proposition and apply it to the dropped-feature example below. Thus $\Psi(u)$ is the exact worst case over candidate representations compatible with the source-error constraint $\overline{A}(h) \le u$. A maximizing representation witnesses that the source-error constraint permits target representation error $\Psi(u)$. An algorithm-specific guarantee can be smaller by using information beyond the value of $\overline{A}(\widehat{h})$.

A useful transfer guarantee proves that $\Psi(u)$ is small in the relevant range. A common bound is

$$\Psi(u) \le \kappa u + \beta, \qquad \kappa, \beta \ge 0. \tag{13.18}$$

The multiplier measures how strongly source accuracy constrains target accuracy. The additive term allows imperfect transfer. The examples in Section 13.3.1.2 bound the envelope by comparing the errors induced by every candidate representation.

13.3.1.1 A source-to-target theorem

With $\widehat{h}$ frozen, target training searches only over the head class. Choose a head with

$$R_{S_0}(\widehat{g}_0 \circ \widehat{h}) \le \min_{g \in \mathcal{G}} R_{S_0}(g \circ \widehat{h}) + \eta_{\mathrm{trg}}, \tag{13.19}$$

where $\eta_{\mathrm{trg}} \ge 0$ and R_{S_0} uses the m target examples.

Theorem 13.3.2 (Transfer from source representation error)**.** *Assume Eq. (13.19). Suppose that, for some $\varepsilon_{\mathrm{src}} \geq 0$ and $0 < \delta_{\mathrm{src}}, \delta_{\mathrm{trg}} < 1$, the source procedure satisfies*

$$\mathbb{P}\{\overline{A}(\widehat{h}) \leq \varepsilon_{\mathrm{src}}\} \geq 1 - \delta_{\mathrm{src}}.$$

Put

$$b_{\mathrm{trg}}(\delta_{\mathrm{trg}}) = \sqrt{\frac{2\ln(2|\mathcal{G}|/\delta_{\mathrm{trg}})}{m}}.$$

With probability at least $1 - \delta_{\mathrm{src}} - \delta_{\mathrm{trg}}$,

$$R_0(\widehat{g}_0 \circ \widehat{h}) - R_0^\star \leq \Psi(\varepsilon_{\mathrm{src}}) + \eta_{\mathrm{trg}} + b_{\mathrm{trg}}(\delta_{\mathrm{trg}}). \tag{13.20}$$

In particular, agnostic joint ERM gives

$$R_0(\widehat{g}_0 \circ \widehat{h}) - R_0^\star \leq \Psi(\Delta_{\mathrm{share}} + \eta_{\mathrm{src}} + b_{\mathrm{src}}(\delta_{\mathrm{src}})) + \eta_{\mathrm{trg}} + b_{\mathrm{trg}}(\delta_{\mathrm{trg}}). \tag{13.21}$$

For an exactly optimized, jointly realizable source problem, one may instead take $\varepsilon_{\mathrm{src}} = a_{\mathrm{src}}(\delta_{\mathrm{src}})$.

Proof. Condition on the source data and source-learner randomization, so that $\widehat{h}$ is fixed and the target sample still has law $P_0^{\otimes m}$. Apply the Hoeffding–union-bound argument to the $|\mathcal{G}|$ predictors $g \circ \widehat{h}$. With conditional probability at least $1 - \delta_{\mathrm{trg}}$,

$$R_0(\widehat{g}_0 \circ \widehat{h}) - \min_{g \in \mathcal{G}} R_0(g \circ \widehat{h}) \leq \eta_{\mathrm{trg}} + b_{\mathrm{trg}}(\delta_{\mathrm{trg}}).$$

The conditional failure probability is at most δ_{trg} for every source realization, so the unconditional failure probability is also at most δ_{trg}. On the intersection with the source event,

$$\min_g R_0(g \circ \widehat{h}) - R_0^\star = A_0(\widehat{h}) \leq \Psi(\overline{A}(\widehat{h})) \leq \Psi(\varepsilon_{\mathrm{src}}).$$

Adding the inequalities proves Eq. (13.20). Use Theorem 13.2.4 or Theorem 13.2.2 for the respective choices of $\varepsilon_{\mathrm{src}}$. In the realizable case, $\overline{A}(\widehat{h}) \leq \overline{R}(\widehat{h}, \widehat{\boldsymbol{g}})$. □

The target fitting cost contains $\ln|\mathcal{G}|$, not $\ln|\mathcal{H}|$. Learning the target from scratch has the finite-class ERM bound

$$R_0(\widehat{f}_{\mathrm{scratch}}) - R_0^\star \leq \sqrt{\frac{2\{\ln|\mathcal{H}| + \ln|\mathcal{G}| + \ln(2/\delta_{\mathrm{trg}})\}}{m}}$$

for an exact ERM. Transfer can improve the guarantee when $\Psi(\varepsilon_{\mathrm{src}})$ is smaller than the estimation cost saved on the target. The target need not be sampled from a distribution over tasks; Eq. (13.17) is the explicit connection to the fixed source tasks.

The theorem also answers the individual-source-task question. Fix $t \in [T]$ and define

$$\Psi_t(u) = \max\big(\{A_t(h) : h \in \mathcal{H},\ \overline{A}(h) \leq u\} \cup \{0\}\big).$$

After freezing $\widehat{h}$ and refitting task t's head on fresh examples, Eq. (13.20) holds for that task with Ψ_t. For q designated tasks, the same source event and head failure probabilities δ_j give simultaneous guarantees with failure probability at most $\delta_{\mathrm{src}} + \sum_{j=1}^q \delta_j$. The guarantee uses refitting samples that are independent of the source examples used to learn $\widehat{h}$. It concerns the newly fitted heads; guarantees for the original fitted heads require corresponding individual-task source bounds.

13.3.1.2 Bounding the transfer envelope

The next two examples identify source tasks that constrain target-relevant features and source tasks that leave a necessary feature unconstrained.

Label-reversed tasks. Let's return to the example of Section 13.2.2.2. In Eq. (13.7), let the target also have a uniform input with the same law μ as the source tasks, and label $h^\star(X) \oplus b_0$. For a candidate h, put

$$q(h) = \mu\big(\{x : h(x) \neq h^\star(x)\}\big).$$

Recall that $\mathcal{G} = \{g_0, g_1\}$ with $g_b(z) = z \oplus b$ for $b \in \{0,1\}$. For each task $t \in [T]$, the target predictor is $g_{b_t} \circ h^\star$. Since each g_b is a bijection on $\{0,1\}$,

$$R_t(g_{b_t} \circ h) = \mu\big(\{x : g_{b_t}(h(x)) \neq g_{b_t}(h^\star(x))\}\big) = q(h), \qquad R_t(g_{1-b_t} \circ h) = 1 - q(h).$$

Since $R_t^* = 0$,

$$A_t(h) = \min_{g \in \mathcal{G}} R_t(g \circ h) = \min\{q(h), 1 - q(h)\} = A_0(h) = \overline{A}(h)\,. \tag{13.22}$$

Thus $\Psi(u) \leq u$, independently of T. The transfer theorem (Theorem 13.3.2)) gives, for exact source and target ERM,

$$R_0(\widehat{g}_0 \circ \widehat{h}) \leq \frac{(d+T)\ln 2 + \ln(1/\delta_{\mathrm{src}})}{nT} + \sqrt{\frac{2\ln(4/\delta_{\mathrm{trg}})}{m}}. \tag{13.23}$$

For fixed accuracy and confidence, T of order d permits constant n and constant m. A learner with only the m target observations still has the expected-error lower bound $\frac{1}{2}(1 - 1/d)^m$ for some target map. For comparison in expectation, a risk bound of ε with failure probability δ implies expected risk at most $\varepsilon + \delta$. Choosing both constants small gives a genuine individual-task advantage.

A target feature unconstrained by the source tasks. Let $\mathcal{X} = \mathcal{Z} = \{0,1\}^2$, let $\mathcal{Y} = \{0,1\}$, and use 0–1 loss. Let U, V be independent fair bits and set $X = (U, V)$. The representation and head classes are

$$\mathcal{H} = \{h_{\mathrm{full}}, h_{\mathrm{drop}}\}, \qquad h_{\mathrm{full}}(u,v) = (u,v), \qquad h_{\mathrm{drop}}(u,v) = (u,0),$$

and

$$\mathcal{G} = \{\pi_1, \pi_2\}, \qquad \pi_1(z_1,z_2) = z_1, \qquad \pi_2(z_1,z_2) = z_2.$$

An observation from each source task is $(X,Y) = ((U,V),U)$, whereas a target observation is $(X,Y) = ((U,V),V)$. Thus the source and target input laws are the same and uniform on $\{0,1\}^2$, while the label functions differ. The full representation has zero risk on every source task using π_1 and zero target risk using π_2, so $R_t^\star = R_0^\star = 0$. For the dropped representation,

$$R_t(\pi_1 \circ h_{\mathrm{drop}}) = 0 \quad (t \in [T]), \qquad R_0(\pi_1 \circ h_{\mathrm{drop}}) = R_0(\pi_2 \circ h_{\mathrm{drop}}) = \frac{1}{2}.$$

Consequently,

$$\overline{A}(h_{\mathrm{drop}}) = 0, \qquad A_0(h_{\mathrm{drop}}) = \frac{1}{2}.$$

No source loss depends on whether a representation retains V. The transfer envelope therefore satisfies $\Psi(0) = 1/2$, and repeating the source tasks leaves the target feature V unconstrained.

The last two examples concern whether the source task objectives constrain the features needed by the target. In the label-reversal example, every source task supplies such a constraint. In the present example, no source task constrains retention of V.

Lecture 7's coverage condition addresses a separate issue: comparison of errors under different input laws. The two forms of coverage can be imposed together. For an intermediate representation error $A_{\mathrm{mid}}(h)$, suppose an input-law comparison gives $A_0(h) \le \kappa_1 A_{\mathrm{mid}}(h) + \beta_1$ and a source-task comparison gives $A_{\mathrm{mid}}(h) \le \kappa_2 \overline{A}(h) + \beta_2$. Substitution gives

$$A_0(h) \le \kappa_1\kappa_2 \overline{A}(h) + \kappa_1\beta_2 + \beta_1.$$

Thus distributional coverage and source-task coverage address different obstacles and combine directly.

13.3.2 Transfer with linear representations

We now give an efficient algorithm for the transfer problem. The learner estimates a linear representation from source tasks by an eigendecomposition, then fits a linear head on target data. The model lets us verify the source-to-target relation and bound the risk of the resulting target predictor.

This is the same representation-learning template as in the earlier finite-class analysis, specialized to linear maps and squared loss. Concretely, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Z} = \mathbb{R}^k$, $\mathcal{Y} = \mathbb{R}$, and we use the squared loss $\ell(y, y') = (y - y')^2$. The representation and head classes are

$$\mathcal{H} = \{h_C : \mathbb{R}^d \to \mathbb{R}^k : h_C(x) = C^\top x,\ C \in \mathbb{R}^{d\times k},\ C^\top C = I_k\}, \tag{13.24}$$

$$\mathcal{G} = \{g_b : \mathbb{R}^k \to \mathbb{R} : g_b(z) = b^\top z,\ b \in \mathbb{R}^k\}. \tag{13.25}$$

The composite predictor is $f_{C,b}(x) = g_b(h_C(x)) = b^\top C^\top x = x^\top C b$. Thus the linear model is the previous class-based setup with explicit, infinite classes $\mathcal{H}$ and $\mathcal{G}$.

13.3.2.1 The learning problem

There are T source tasks and one target task, indexed by 0. Let $1 \le k < d$. The unknown shared object is a k-dimensional subspace of $\mathbb{R}^d$. Choose a matrix $B \in \mathbb{R}^{d\times k}$ whose orthonormal columns form a basis for this subspace, and let $a_0, a_1, \ldots, a_T \in \mathbb{R}^k$ be the task coordinates. The orthonormal convention preserves every full-column-rank linear model: if a feature matrix factors as $M = BR$, replacing a task coefficient c_t by $a_t = Rc_t$ gives $Mc_t = Ba_t$. An observation from task t satisfies

$$Y = X^\top B a_t + \varepsilon, \qquad X \sim \mathcal{N}(0, I_d), \tag{13.26}$$

where the noise has mean zero and is independent of X, with finite variance σ^2 on each source task and σ_0^2 on the target task. The learner is given k, samples $S_1, \ldots, S_T$ of n examples each, and a target sample S_0 of m examples. All covariates and noises in these samples are mutually independent. Write $N = nT$ for the source sample size.

Thus each C selects a candidate k-dimensional feature subspace and an orthonormal basis for it. Because $\mathcal{G}$ contains every linear head, a change of basis changes the head coordinates but not the available composite predictors. The normalization also makes $CC^\top$ the orthogonal projector onto the candidate subspace, as used below. Use R_t, $R_t^\star$, and A_t for these classes and squared loss. The learner estimates $\widehat{B}$ from the source samples alone and fits $\widehat{a}_0$ on S_0, returning $(h_{\widehat{B}}, g_{\widehat{a}_0})$. The

predictor is $\widehat{f}_0(x) = x^\top \widehat{\theta}$, where $\widehat{\theta} = \widehat{B}\widehat{a}_0$. The objective is small expected target excess risk, which in this model is

$$\mathbb{E}[R_0(\widehat{f}_0) - R_0^\star] = \mathbb{E}[\|\widehat{\theta} - Ba_0\|_2^2], \qquad R_0^\star = \sigma_0^2. \tag{13.27}$$

The statistical point is the same as in the finite-class transfer analysis: source training constrains the representation only insofar as it rules out bad shared subspaces. In this linear model the constraint is explicit. For a candidate representation C, the squared norm of the component of Ba_0 orthogonal to $\mathrm{span}(C)$ is the target representation error $A_0(C)$. The comparison is therefore between the target representation error $A_0(C)$ and the average source representation error $\overline{A}(C)$, and the source-task matrix $D = \frac{1}{T}\sum_{t=1}^T a_t a_t^\top$ measures how strongly the source tasks cover the relevant directions. For well-spread sources, $D \succ 0$ and a target direction with small source representation error inherits a comparable target representation error. This is the linear analogue of the earlier calibration bound, with the quantitative factor $\kappa_0 = a_0^\top D^{-1} a_0$.

The model gives two distinct comparisons with the earlier theory. The statistical comparison concerns the number of target examples needed after the source tasks have identified a useful representation. The corollary below bounds the expected target excess risk by a source term of order $\|a_0\|_2^2 V/(nT\lambda^2)$ and a target head-fitting term $\sigma_0^2 k/(m-k-1)$. Target-only least squares instead has target term $\sigma_0^2 d/(m-d-1)$. Thus the ideal sharing benefit is a reduction from ambient dimension d to representation dimension k in the target-label cost; the source cost is amortized across the T source tasks. The displayed bound also quantifies the price of imperfect subspace recovery. This conclusion is a property of the stated Gaussian linear model.

The algorithmic comparison asks how to obtain the shared representation. Joint source ERM is a natural benchmark:

$$\min_{C^\top C = I_k,\, b_1,\dots,b_T} \frac{1}{N}\sum_{t=1}^T \sum_{j=1}^n \big(Y_{tj} - b_t^\top C^\top X_{tj}\big)^2.$$

This objective is precisely the source-training objective for a two-layer linear network with a shared first layer $C^\top$ and task-specific output weights b_t. It is a finite-dimensional rank-constrained least-squares problem. At the population level, minimizing over the heads makes its excess objective

$$\mathrm{tr}\Big((I_d - CC^\top)BDB^\top(I_d - CC^\top)\Big).$$

If $D \succ 0$, the global minimizers therefore satisfy $\mathrm{span}(C) = \mathrm{span}(B)$. Thus population ERM identifies the desired representation. With finite samples, one can instead train the factorized empirical objective by a first-order method, or minimize over each b_t in closed form and optimize the resulting profile objective over C. Establishing that either procedure reaches a statistically accurate solution requires an optimization-landscape analysis in addition to a generalization argument; related results are described in the bibliographical notes.

Instead, the linear-Gaussian model supplies a more specific route. The population second moment of $Y^2 XX^\top$ has the shared subspace as its leading k-dimensional eigenspace. We estimate that moment and compute its leading eigenvectors. The spectral estimator is therefore a deliberate choice: it is an explicit polynomial-time procedure whose source-data dependence and subspace error can be analyzed directly. The following calculations show both why the source tasks identify the joint representation and how the statistical saving survives after fitting the target head.

13.3.2.2 Source diversity controls target representation error

We first ask which source tasks constrain the representation needed by the target. The Gaussian input model makes representation error a projection error, so the comparison reduces to a relation between the task coefficients.

For $u = Cb - Ba_t$, the prediction residual is $X^\top u - \varepsilon$. The model assumptions imply

$$\mathbb{E}[XX^\top] = I_d, \qquad \mathbb{E}[\varepsilon X] = \mathbb{E}[\varepsilon]\mathbb{E}[X] = 0,$$

where the second identity uses the independence of ε and X. Consequently,

$$R_t(g_b \circ h_C) = \mathbb{E}[(X^\top u - \varepsilon)^2] = u^\top \mathbb{E}[XX^\top]u - 2u^\top \mathbb{E}[\varepsilon X] + \mathbb{E}[\varepsilon^2] = \|Cb - Ba_t\|_2^2 + \mathbb{E}[\varepsilon^2].$$

Since $C = B$ is allowed, $R_t^\star = \mathbb{E}[\varepsilon^2]$. Minimizing over b gives

$$A_t(C) = \|(I_d - CC^\top)Ba_t\|_2^2. \tag{13.28}$$

Define

$$D = \frac{1}{T}\sum_{t=1}^{T} a_t a_t^\top, \qquad E_C = (I_d - CC^\top)B. \tag{13.29}$$

Then $\overline{A}(C) = \operatorname{tr}(E_C D E_C^\top)$.

Proposition 13.3.3 (Linear source-to-target comparison)**.** *If $D \succ 0$, every candidate C satisfies*

$$A_0(C) \le \kappa_0\, \overline{A}(C), \qquad \kappa_0 = a_0^\top D^{-1} a_0 \le \frac{\|a_0\|_2^2}{\lambda_{\min}(D)}. \tag{13.30}$$

Proof. Cauchy–Schwarz applied to $D^{-1/2}a_0$ and $D^{1/2}v$ gives $(v^\top a_0)^2 \le (a_0^\top D^{-1} a_0) v^\top D v$ for every v. Apply this inequality to each row of E_C and add: $\|E_C a_0\|_2^2 \le \kappa_0 \operatorname{tr}(E_C D E_C^\top)$. Finally, $D \succeq \lambda_{\min}(D) I_k$ implies $D^{-1} \preceq \lambda_{\min}(D)^{-1} I_k$. □

The factor depends on the target direction as well as the source tasks. A target aligned with a well-covered direction can have a small factor even when other directions are poorly covered.[2] If $\|a_0\|_2 \le L$, the bound $\Psi(u) \le L^2 u/\lambda_{\min}(D)$ holds simultaneously for all such targets. The matrix D gives a concrete measure of how strongly the source tasks constrain the shared representation.

13.3.2.3 Estimating the representation from source data

We estimate the shared subspace using a second moment of the source observations. The resulting *spectral estimator* forms a representation from leading eigenvectors of a data-dependent matrix. For the sample bound below, additionally assume $\varepsilon_{tj} \sim \mathcal{N}(0, \sigma^2)$ on the source tasks. The moment identity itself only uses centered noise of variance σ^2 independent of the covariates.

The estimator belongs to the spectral family of representation-learning methods. In linear models, the representation is a subspace and the estimator recovers that subspace as the leading eigenspace of a covariance or moment matrix. Kernel PCA, Laplacian eigenmaps, and spectral clustering replace the linear moment matrix by a kernel matrix, graph Laplacian, or population operator. Their coordinates can depend nonlinearly on the original input, while the learned representation remains an eigenspace or a collection of eigenfunctions. Spectral methods therefore cover a broad low-rank and operator-based family. The bibliographical notes place this family among other representation-learning objectives.

Proposition 13.3.4 (Gaussian task moment)**.** *For one observation from a task with parameter Ba,*

$$\mathbb{E}[Y^2 XX^\top] = (\|a\|_2^2 + \sigma^2) I_d + 2Baa^\top B^\top. \tag{13.31}$$

Consequently the average population moment is

$$M_\star = (\operatorname{tr} D + \sigma^2) I_d + 2BDB^\top. \tag{13.32}$$

If $D \succ 0$, its leading k-dimensional eigenspace is $\operatorname{span}(B)$*, with eigengap* $2\lambda_{\min}(D)$*.*

Proof. Put $u = Ba$. Since ε is independent of X and has mean zero,

$$\mathbb{E}[2\varepsilon(X^\top u)XX^\top] = 2\mathbb{E}[\varepsilon]\mathbb{E}[(X^\top u)XX^\top] = 0, \qquad \mathbb{E}[\varepsilon^2 XX^\top] = \mathbb{E}[\varepsilon^2]\mathbb{E}[XX^\top] = \sigma^2 I_d.$$

For standard Gaussian coordinates,

$$\mathbb{E}[X_i X_j X_r X_s] = \mathbf{1}_{\{i=j\}}\mathbf{1}_{\{r=s\}} + \mathbf{1}_{\{i=r\}}\mathbf{1}_{\{j=s\}} + \mathbf{1}_{\{i=s\}}\mathbf{1}_{\{j=r\}}.$$

Expanding $(X^\top u)^2$ therefore gives $\mathbb{E}[(X^\top u)^2 XX^\top] = \|u\|_2^2 I_d + 2uu^\top$. Since $u = Ba$ and $B^\top B = I_k$,

$$\|u\|_2^2 = a^\top B^\top Ba = \|a\|_2^2, \qquad uu^\top = Baa^\top B^\top.$$

Combining the three terms in $Y^2 = (X^\top u + \varepsilon)^2$ now proves Eq. (13.31). Average this identity over the source tasks:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Y_{t1}^2 X_{t1}X_{t1}^\top] &= \left(\frac{1}{T}\sum_{t=1}^{T}\|a_t\|_2^2 + \sigma^2\right) I_d + 2B\left(\frac{1}{T}\sum_{t=1}^{T} a_t a_t^\top\right) B^\top \\ &= (\operatorname{tr} D + \sigma^2) I_d + 2BDB^\top,\end{aligned}$$

because $T^{-1}\sum_t \|a_t\|_2^2 = \operatorname{tr} D$. This proves Eq. (13.32). Put $c = \operatorname{tr} D + \sigma^2$. If $Dq = \lambda_i(D)q$, then

$$M_\star(Bq) = cBq + 2BDq = (c + 2\lambda_i(D))Bq.$$

For $v \perp \operatorname{span}(B)$, the identity $B^\top v = 0$ gives $M_\star v = cv$. Thus the k eigenvalues on $\operatorname{span}(B)$ are $c + 2\lambda_i(D)$, every eigenvalue on its orthogonal complement is c, and the eigengap is $2\lambda_{\min}(D)$. □

The estimator forms

$$\widehat{M} = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n} Y_{tj}^2 X_{tj}X_{tj}^\top \tag{13.33}$$

and takes its leading k orthonormal eigenvectors as the columns of $\widehat{B}$. Computing the moment and a dense eigendecomposition costs $O(Nd^2 + d^3)$ arithmetic operations. The procedure consists of one matrix accumulation followed by one eigendecomposition.

Here is an elementary finite-sample guarantee that makes the statistical and computational claim explicit. Sharper analyses improve its dimension dependence.

Davis–Kahan bound used below. Let M and $\widehat{M} = M + F$ be symmetric matrices. Suppose the leading k-dimensional eigenspace of M has orthonormal basis B and eigengap

$$\gamma = \lambda_k(M) - \lambda_{k+1}(M) > 0.$$

If $\widehat{B}$ contains leading k orthonormal eigenvectors of $\widehat{M}$, then the Davis–Kahan theorem gives

$$\|(I_d - \widehat{B}\widehat{B}^\top)B\|_{\mathrm{op}} \le \min\left\{1, \frac{2\|F\|_{\mathrm{op}}}{\gamma}\right\}. \tag{13.34}$$

The left side is the sine of the largest principal angle between the two eigenspaces.

Theorem 13.3.5 (A finite-sample spectral guarantee)**.** *Suppose* $\lambda = \lambda_{\min}(D) > 0$*, and let*

$$V = \max_{t\in[T]} \mathbb{E}[Y_{t1}^4 \|X_{t1}\|_2^4].$$

The spectral estimator satisfies

$$\mathbb{E}[A_0(\widehat{B})] \le \frac{\|a_0\|_2^2 V}{N\lambda^2}. \tag{13.35}$$

In the Gaussian model, if $s^2 = \max_t(\|a_t\|_2^2 + \sigma^2)$*, then*

$$V \le \sqrt{105}\, s^4 \sqrt{d(d+2)(d+4)(d+6)}.$$

Proof. The goal is to control the target representation error $\mathbb{E}[A_0(\widehat{B})]$. The representation-error identity Eq. (13.28), which also starts the proof of Proposition 13.3.3, gives

$$\begin{aligned} A_0(\widehat{B}) &= \|(I_d - \widehat{B}\widehat{B}^\top)Ba_0\|_2^2 \\ &\le \|a_0\|_2^2 \|(I_d - \widehat{B}\widehat{B}^\top)B\|_{\mathrm{op}}^2. \end{aligned} \tag{13.36}$$

Thus it suffices to control the angle between $\mathrm{span}(\widehat{B})$ and $\mathrm{span}(B)$.

Proposition 13.3.4 identifies $\mathrm{span}(B)$ as the leading eigenspace of $M_\star$ and gives the population eigengap

$$\gamma = 2\lambda, \qquad \lambda = \lambda_{\min}(D).$$

Put $F = \widehat{M} - M_\star$. Applying Eq. (13.34) with this value of γ gives

$$\|(I_d - \widehat{B}\widehat{B}^\top)B\|_{\mathrm{op}} \le \min\left\{1, \frac{\|F\|_{\mathrm{op}}}{\lambda}\right\} \le \frac{\|F\|_{\mathrm{F}}}{\lambda}.$$

Substitution into Eq. (13.36) yields

$$A_0(\widehat{B}) \le \frac{\|a_0\|_2^2}{\lambda^2}\|F\|_{\mathrm{F}}^2. \tag{13.37}$$

The remaining task is therefore to prove $\mathbb{E}[\|F\|_{\mathrm{F}}^2] \le V/N$.

For this purpose, define

$$W_{tj} = Y_{tj}^2 X_{tj} X_{tj}^\top, \qquad Q_{tj} = W_{tj} - \mathbb{E}[W_{tj}].$$

Proposition 13.3.4 and $N = nT$ give

$$\mathbb{E}[\widehat{M}] = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{n}\mathbb{E}[W_{tj}] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_{t1}] = M_\star.$$

Consequently, $F = N^{-1}\sum_{t,j} Q_{tj}$. The matrices Q_{tj} are independent and satisfy $\mathbb{E}[Q_{tj}] = 0$. Hence, for $(t,j) \ne (s,r)$,

$$\mathbb{E}[\langle Q_{tj}, Q_{sr}\rangle_{\mathrm{F}}] = \langle \mathbb{E}[Q_{tj}], \mathbb{E}[Q_{sr}]\rangle_{\mathrm{F}} = 0.$$

Expanding the squared Frobenius norm therefore gives

$$\begin{aligned} \mathbb{E}[\|F\|_{\mathrm{F}}^2] &= \frac{1}{N^2}\sum_{t,j}\mathbb{E}[\|Q_{tj}\|_{\mathrm{F}}^2] && \le \frac{1}{N^2}\sum_{t,j}\mathbb{E}[\|W_{tj}\|_{\mathrm{F}}^2] \\ &= \frac{1}{N^2}\sum_{t,j}\mathbb{E}[Y_{tj}^4\|X_{tj}X_{tj}^\top\|_{\mathrm{F}}^2] \\ &= \frac{1}{N^2}\sum_{t,j}\mathbb{E}[Y_{tj}^4\|X_{tj}\|_2^4] \le \frac{V}{N}. \end{aligned}$$

The inequality uses $\mathbb{E}[\|W - \mathbb{E}[W]\|_{\mathrm{F}}^2] \le \mathbb{E}[\|W\|_{\mathrm{F}}^2]$, and the penultimate identity uses $\|xx^\top\|_{\mathrm{F}}^2 = \mathrm{tr}(xx^\top xx^\top) = \|x\|_2^4$. Taking expectations in Eq. (13.37) now proves Eq. (13.35).

It remains to bound V under Gaussian noise. For each source task, $Y_{t1} = X_{t1}^\top B a_t + \varepsilon_{t1}$ is centered Gaussian with variance $\|a_t\|_2^2 + \sigma^2 \le s^2$. Hence $\mathbb{E}[Y_{t1}^8] \le 105 s^8$. Cauchy–Schwarz gives $\mathbb{E}[Y_{t1}^4 \|X_{t1}\|_2^4] \le \sqrt{\mathbb{E}[Y_{t1}^8]\mathbb{E}[\|X_{t1}\|_2^8]}$. Since $\mathbb{E}[\|X_{t1}\|_2^8] = d(d+2)(d+4)(d+6)$, we obtain

$$\mathbb{E}[Y_{t1}^4 \|X_{t1}\|_2^4] \le \sqrt{105}\, s^4 \sqrt{d(d+2)(d+4)(d+6)}.$$

Taking the maximum over t proves the stated bound on V. □

13.3.2.4 Fitting the target head and bounding prediction risk

With $\widehat{B}$ fixed, target training is ordinary least squares in k features. Let $Z_0 \in \mathbb{R}^{m \times k}$ have rows $(\widehat{B}^\top X_{0j})^\top$ and let $\boldsymbol{Y}_0 = (Y_{01}, \dots, Y_{0m})^\top$. For $m > k+1$, set

$$\widehat{a}_0 = (Z_0^\top Z_0)^{-1} Z_0^\top \boldsymbol{Y}_0, \qquad \widehat{\theta} = \widehat{B}\widehat{a}_0. \tag{13.38}$$

Computing the target features and fitting the head costs $O(mdk + mk^2 + k^3)$ arithmetic operations. The exact Gaussian calculation in Appendix 13.A converts the representation bound into a guarantee for the final target predictor.

Corollary 13.3.6 (Linear transfer guarantee)**.** *Under the assumptions of Theorem 13.3.5, let $\widehat{B}$ be the spectral estimate and fit the target head by Eq. (13.38) using $m > k+1$ target examples. With λ and V as in that theorem, the expected target excess risk is bounded by*

$$\mathbb{E}[R_0(\widehat{f}_0) - R_0^\star] \le \left(1 + \frac{k}{m-k-1}\right) \frac{\|a_0\|_2^2 V}{N\lambda^2} + \sigma_0^2 \frac{k}{m-k-1}. \tag{13.39}$$

Proof. The target sample is independent of $\widehat{B}$. Conditional on $\widehat{B}$, Theorem 13.A.1 gives

$$\mathbb{E}[\|\widehat{\theta} - B a_0\|_2^2 \mid \widehat{B}] = \left(1 + \frac{k}{m-k-1}\right) A_0(\widehat{B}) + \sigma_0^2 \frac{k}{m-k-1}.$$

Average over the source data, apply Theorem 13.3.5, and use Eq. (13.27). □

The first term accounts for learning the representation from source data; the second accounts for fitting the head from target data. For comparison, ordinary least squares using only the target sample in $\mathbb{R}^d$ has expected excess risk $\sigma_0^2 d/(m-d-1)$ when $m > d+1$. Accurate source-based features reduce the dimension in the target estimation term from d to k. Together with the arithmetic costs above, Corollary 13.3.6 gives a computational and statistical guarantee for transfer.

Connection to neural representations. A downstream task can use a neural network's frozen features and a newly fitted head. The transfer analyses explain what a guarantee for that use requires: source training must control the target representation error, and target data must suffice to fit the head. For language-model pretraining, the remaining question is how next-token loss constrains downstream representation error.

13.4 Take-home messages

- Joint ERM pays the representation complexity once across the source tasks and the head complexity once per task. The shared boxes example gives a strict statistical advantage over independent learning.
- Misspecification contributes the sharing cost Δ_{share}. Reduced estimation error must compensate for any added approximation error.
- A bound on $A_0(h)$ in terms of $\overline{A}(h)$ converts source learning into a guarantee for a particular target. The examples verify the relation and exhibit a case where it fails.
- With independent target examples and a frozen representation, head fitting pays only the head-class complexity. Simultaneous individual-task guarantees require suitable source-to-task relations for those tasks.
- In the Gaussian linear model, a spectral representation estimate followed by target least squares has an explicit computation cost and a finite-sample target-risk guarantee.

Endnotes

1. We choose the multi-task setting of Section 13.2.1 where the task identities are known and different heads needs to be learnt for each task for pedagogical reasons. An alternative, and in many ways more common scenario, is when the data is not grouped based on tasks (thus the information of which task a given example belongs to is hidden) and/or the same head is to be used across all the tasks. Examples abound; one obvious case is language modeling with multiple languages (e.g., English, German, Hungarian, Chinese, Python, Java, ...). In these domain-agnostic (or task-agnostic) setting, the representation learnt could separate the language agnostic content from the language identity, thus making the learning of the head across many different languages more efficient.

2. Positive definiteness of D is sufficient for all targets. For a fixed target with $a_0 \in \text{range}(D)$, the same proof on that range gives $\kappa_0 = a_0^\top D^\dagger a_0$, where $D^\dagger$ is the Moore–Penrose inverse. Coverage of directions unused by the target is unnecessary.

Bibliographical notes

For a broad introduction, Bengio et al. [2013] survey representation learning, and Chapter 15 of Goodfellow et al. [2016] gives a textbook treatment. The following map organizes several major families by the information used to learn the representation and by the property retained in the representation.

Linear, latent, and spectral structure. PCA retains directions of large variance; factor models and ICA seek latent coordinates with additional statistical structure; graph and kernel methods use eigenvectors or eigenfunctions of similarity operators. The review by Jolliffe and Cadima [2016], the ICA book by Hyvärinen et al. [2001], and the spectral clustering tutorial by von Luxburg [2007] provide readable introductions. Belkin and Niyogi [2003] show how a graph Laplacian produces nonlinear coordinates for data near a manifold.

Supervised task sharing and transfer. Labels from several tasks train features that support task-specific heads. Multitask learning, learning-to-learn, and the linear model in this lecture belong to this family. The statistical sources most closely connected to the lecture are discussed below.

Reconstruction and generative modeling. Autoencoders, probabilistic latent-variable models, and related generative methods learn coordinates that retain information needed to reconstruct or model the observations. Bengio et al. [2013] and Goodfellow et al. [2016] connect these objectives to deep representation learning.

Predictive and contrastive self-supervision. Context prediction, masked or next-token prediction, and contrastive objectives derive training signals from the observations themselves. Le-Khac et al. [2020] give an organized review of the contrastive branch.

Representations for sequential decisions. Control and reinforcement-learning methods seek state variables that preserve reward-relevant and transition-relevant information. The overview by Lesort et al. [2018] organizes reconstruction, forward-model, inverse-model, and robotic-prior approaches.

Geometric and causal structure. Equivariant methods encode symmetries and relational structure; causal representation learning seeks variables that support interventions and distributional changes. Bronstein et al. [2021] survey geometric deep learning, and Schölkopf et al. [2021] survey causal representation learning.

Multitask feature learning and transfer have a substantial statistical literature for general classes of representations. Baxter [2000] studies learning a hypothesis space from tasks sampled from an environment. Maurer et al. [2016] gives general representation-class bounds for both observed-task prediction and learning new tasks, with separate costs for representations and task-specific predictors. The finite-class results here are elementary versions of this complexity separation, proved using the tools already developed in the course.

The closest source for bounds such as Eq. (13.18) is the task-diversity condition of Tripuraneni et al. [2020], Definition 3. Their condition bounds a worst-case target representation discrepancy by an average source discrepancy divided by a diversity parameter, plus an additive error. Our fixed-target formulation uses the transfer envelope Ψ and class-relative approximation errors; the displayed finite-class theorem is proved directly. The transfer envelope Ψ expresses the connection to the fixed source tasks directly.

The Gaussian linear model and moment estimator follow Tripuraneni et al. [2021]. The simple Frobenius-moment bound in Theorem 13.3.5 is deliberately weaker than sharp analyses of this model and provides a complete finite-sample argument. The eigenspace perturbation step is the Davis–Kahan theorem [Davis and Kahan, 1970]. The inverse-Gram expectation in Eq. (13.45) is the standard inverse-Wishart mean identity.

Tripuraneni et al. [2021] also study optimization-based representation learning. They use a Burer–Monteiro factorization of the task-parameter matrix and add a balancing regularizer. Under their sub-Gaussian design, task-diversity, incoherence, sample-size, and uniform-task-sampling conditions, every local minimum in a specified interior region recovers the shared subspace up to statistical error. Their local-minimum result, together with general results on escaping strict saddles, gives a route from first-order optimization to an accurate representation. Their theorem applies to the regularized factorization in its stated constraint region.

Task-boundary terminology is illustrated by Aljundi et al. [2019]. The optional intrinsic-dimension comparison in Appendix 13.B uses Kohler et al. [2026], Kpotufe [2011], Kpotufe and Garg [2013], and Farahmand et al. [2007].

Glossary

Risk notation A sample in the subscript marks an empirical risk, so R_t uses the law P_t and R_{S_t} uses the sample S_t. A bar averages over the T source tasks, giving $\overline{R}$ and $\overline{R}_S$. A hat marks a quantity returned by a learner, such as $\widehat{h}$ or $\widehat{B}$. The subscripts src and trg distinguish source- and target-side quantities.

Representation-learning algorithm A map $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H} \times \mathcal{G}$ returning a representation and a head for supervised prediction.

Representation and head Maps $h : \mathcal{X} \to \mathcal{Z}$ and $g : \mathcal{Z} \to \mathcal{Y}$, whose composition predicts a label. In the multitask setting, h is shared and each task has its own head.

Multitask learning Jointly learning predictors for several tasks and evaluating on those tasks.

Transfer learning Using information learned from source tasks to learn a predictor for a target task; here, fitting a new head on a frozen source-trained representation.

Learning-to-learn / meta-learning Learning how to fit new tasks from experience on source tasks.

Joint ERM Minimization of average empirical task risk over one shared representation and one head for each task.

Representation error $A_t(h)$ The extra population risk imposed on task t by using h, after optimizing the head, relative to the best pair in $\mathcal{H} \times \mathcal{G}$.

Sharing cost Δ_{share} The minimum average representation error over one shared h.

Transfer envelope Ψ The largest target representation error among candidate representations whose average source representation error is at most a specified value.

Task-diversity matrix The matrix $D = T^{-1} \sum_t a_t a_t^\top$ in the linear model. Its quadratic form measures source coverage of a coefficient direction.

Spectral representation estimator An estimator that uses leading eigenvectors of a data-dependent matrix to define the feature map.

Negative transfer The excess risk of a sharing or transfer procedure relative to a specified nonsharing baseline.

Intrinsic dimension A dimension describing the geometry of an input distribution, rather than the complexity of a shared family of tasks.

13.5 Exercises

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. ◎ marks the intended learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Joint learning and transfer



Exercise 1: Finite-class zero-loss bound.

◎ Extend the finite-class consistency argument from classification to bounded losses and independent observations with different laws.

Use the setup immediately preceding Proposition 13.2.1.

(a) Let Z be a random variable with values in $[0,1]$. Prove

$$\mathbb{P}(Z=0) \le 1 - \mathbb{E}[Z].$$

Show that equality holds when Z takes values in $\{0,1\}$.

(b) For a fixed $f \in \mathcal{F}$, use independence and part (a) to prove

$$\mathbb{P}\big(\ell(f(U_i),V_i)=0 \text{ for every } i \in [m]\big) \le e^{-m\overline{R}_m(f)}.$$

(c) Use a union bound to prove the probability statement in Proposition 13.2.1. Your argument must apply to every choice among zero-training-loss members of $\mathcal{F}$, including a randomized choice.

(d) Integrate the resulting tail bound to prove the expected-risk statement. Explain where the assumption that some $f^\star \in \mathcal{F}$ has $R_i(f^\star)=0$ for every i is used.



Exercise 2: The approximation cost of sharing.

◎ Compute a sharing cost when every task is individually realizable, and compare its effect with the estimation cost.

Let U, V be independent fair bits and let $X=(U,V)$ on every task. A fraction p of the T source tasks have label U, and the remaining fraction have label V, where pT is an integer. Take $\mathcal{H}=\{h_U,h_V\}$ with $h_U(U,V)=U$ and $h_V(U,V)=V$, $\mathcal{G}=\{z\mapsto z,\ z\mapsto 1-z\}$, and 0–1 loss. Each task supplies n independent examples, and the task samples are independent.

(a) Compute $R_t^\star$, $A_t(h_U)$, and $A_t(h_V)$ for both kinds of task. Deduce the exact value of Δ_{share}.

(b) Specialize the agnostic joint-ERM bound to these classes, assuming exact empirical optimization.

(c) Suppose $0 < p < 1$. Explain why increasing n cannot make the average risk of a shared predictor tend to zero. Compare with consistent independent learning. Identify the assumption of the realizable joint zero-loss theorem that fails.

Exercise 3: Sharpness of the transfer envelope.

Show that the transfer envelope is the strongest target guarantee implied only by an average source representation-error bound.

Let Ψ be the transfer envelope from Eq. (13.16).

(a) Prove that Ψ is nondecreasing and that $A_0(h) \le \Psi(\overline{A}(h))$ for every $h \in \mathcal{H}$.

(b) Let $\phi : [0, \infty) \to [0, \infty)$ be nondecreasing and satisfy $A_0(h) \le \phi(\overline{A}(h))$ for every $h \in \mathcal{H}$. Prove that $\Psi(u) \le \phi(u)$ for every $u \ge 0$. When $\{h : \overline{A}(h) \le u\}$ is nonempty, prove that some feasible h_u attains $A_0(h_u) = \Psi(u)$. Explain the precise worst-case statement established by these facts.

(c) In the target-feature example of Section 13.3.1.2, compute $\Psi(0)$ and identify a maximizing representation. Give a joint source-ERM tie-breaking rule that always selects this representation. Explain why the value of $\Psi(0)$ does not imply that every source learner selects it.

Exercise 4: Coverage of a target and simultaneous task guarantees.

Turn a quantitative source-to-target relation into useful guarantees for designated individual tasks.

Use the finite-class bounded-loss setup of the lecture, and suppose $\mathbb{P}\{\overline{A}(\widehat{h}) \le \varepsilon_{\mathrm{src}}\} \ge 1 - \delta_{\mathrm{src}}$.

(a) Suppose at least ρT source tasks have the same law as a fixed target P_0, where $0 < \rho \le 1$. Prove that $A_0(h) \le \overline{A}(h)/\rho$ for every h. Explain why the factor depends on the fraction of relevant source tasks, rather than their number alone.

(b) More generally, for targets $j = 1, \dots, q$, suppose known constants $\kappa_j, \beta_j \ge 0$ satisfy $A_j(h) \le \kappa_j \overline{A}(h) + \beta_j$ for all h. Each target supplies m_j fresh i.i.d. examples independent of the source data. Freeze $\widehat{h}$ and fit each head by exact ERM. Give a simultaneous bound on all q excess risks with probability at least $1 - \delta_{\mathrm{src}} - \delta_{\mathrm{heads}}$, assigning each head the same failure probability. Are the target datasets required to be independent of one another for your proof?

Exercise 5: A small target sample can select the head.

Sharpen the generic target-ERM bound when only the orientation of a shared Boolean feature remains to be learned.

Use the label-reversal model of Eq. (13.7). Suppose the source procedure has an event of probability at least $1 - \delta_{\mathrm{src}}$ on which $r := A_0(\widehat{h}) \le \varepsilon \le 1/4$. Freeze $\widehat{h}$ and select one of the two heads by ERM on m fresh target examples independent of all source data.

(a) Conditional on a source sample in that event, prove that the probability of choosing the higher-risk head is at most $\exp(-m/8)$. Allow arbitrary tie breaking.

(b) Deduce a guarantee that the final target risk is at most ε with probability at least $1 - \delta_{\mathrm{src}} - \delta_{\mathrm{trg}}$. Give sufficient n and m when the source learner is joint ERM with n examples from each of T tasks.

(c) For fixed $\varepsilon \in (0, 1/4)$ and fixed confidence, compare the target sample requirement with the boxes lower bound for a learner using only target data. Explain which information the source data supply and which information the target data supply.

HINTS For each target example, the two heads' mistake indicators sum to one. Apply Hoeffding's inequality to the empirical risk of the better head.

Linear representations

Exercise 6: Many repeated tasks can hide a shared feature.

Distinguish the number of source tasks from coverage of the shared feature directions.

Let $d = 3$, $k = 2$, and let every source-task head have the form $a_i = (c_i, 0)^\top$, where $c_i \in \mathbb{R}$. Consider the two feature matrices

$$B = [e_1\ e_2] \qquad \text{and} \qquad \widetilde{B} = [e_1\ e_3].$$

For either feature matrix, source task i generates $Y = X^\top B a_i + \varepsilon$, where $X \sim \mathcal{N}(0, I_3)$ and the same independent noise law is used in both models.

(a) Prove that the joint law of (X, Y) for every source task is identical under B and $\widetilde{B}$.

(b) For uniform sampling from M source tasks, compute the task-diversity matrix

$$D = \frac{1}{M} \sum_{i=1}^{M} a_i a_i^\top$$

and its eigenvalues.

(c) Let the new-task head be $a_{\text{new}} = (0, 1)^\top$. Compute its ambient parameter under each feature matrix. Explain why no procedure using only the source data can determine which of these two new-task parameters is correct.

Exercise 7: The angle between tasks controls the eigengap.

Relate geometric diversity among two task heads to stability of population moment recovery.

Let $k = 2$, and sample two source tasks uniformly. Their unit-norm heads are

$$a_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad a_2 = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}, \qquad 0 < \alpha \le \frac{\pi}{2}.$$

(a) Compute $D = (a_1 a_1^\top + a_2 a_2^\top)/2$ and show that its eigenvalues are

$$\lambda_{\max}(D) = \frac{1 + \cos\alpha}{2}, \qquad \lambda_{\min}(D) = \frac{1 - \cos\alpha}{2}.$$

(b) Under the Gaussian model of Proposition 13.3.4, show that the gap between the second and third eigenvalues of $M_\star$ is $1 - \cos\alpha$.

(c) Determine the leading-order behavior of this eigengap as $\alpha \downarrow 0$. Explain why collecting more observations from two nearly parallel tasks is statistically different from observing two well-separated tasks.

HINTS The trace and determinant of D determine its two eigenvalues. Use $1 - \cos\alpha \sim \alpha^2/2$ for the last part.

Exercise 8: Projection error is representation bias.

Derive the deterministic approximation term that enters the new-task transfer risk.

Let $B, \widehat{B} \in \mathbb{R}^{d\times k}$ have orthonormal columns, and define $P_B = BB^\top$, $\widehat{P} = \widehat{B}\widehat{B}^\top$, and

$$\delta = \|(I_d - \widehat{P})B\|_{\mathrm{op}}.$$

Fix an ambient task parameter $\theta \in \mathbb{R}^d$.

(a) Show that

$$\min_{b\in\mathbb{R}^k} \|\widehat{B}b - \theta\|_2^2 = \|(I_d - \widehat{P})\theta\|_2^2,$$

and identify a minimizing vector b.

(b) If $\theta = Ba$, prove that

$$\|(I_d - \widehat{P})\theta\|_2^2 \le \delta^2\|a\|_2^2.$$

(c) If $\theta = Ba + r$ with $B^\top r = 0$, prove that

$$\|(I_d - \widehat{P})\theta\|_2 \le \delta\|a\|_2 + \|r\|_2.$$

Identify which term is caused by estimating the shared subspace and which term is present even when $\widehat{B} = B$.

Exercise 9: An exact negative-transfer threshold.

Determine when misspecification outweighs the variance saved by a known low-dimensional representation.

Assume the Gaussian new-task model of Theorem 13.A.1. Let $n > d+1$, let $k < d$, and suppose the noise variance is $\sigma^2 > 0$ and the parameter is $\theta = Ba + r$, where $B^\top B = I_k$, $B^\top r = 0$, and the known representation B is used for transfer. Compare ordinary least squares in $\mathrm{span}(B)$ with ordinary least squares in $\mathbb{R}^d$.

(a) Use Theorem 13.A.1 to write the exact expected squared parameter error of the representation-restricted estimator and of the ambient estimator.

(b) Prove that the representation-restricted estimator has smaller expected error exactly when

$$\|r\|_2^2 < \frac{\sigma^2(d-k)}{n-d-1}.$$

(c) Fix $r \neq 0$. Explain how the comparison changes as n grows. Also explain what happens to ambient least squares when $k+1 < n < d$ and when $n \in \{d, d+1\}$. For the latter case, you may use the fact that the expected trace of the inverse Gaussian Gram matrix is infinite.

13.A Exact Gaussian head-estimation risk

This appendix gives the exact head-estimation term used in Eq. (13.39) and a concrete negative-transfer comparison. Fix $\widehat{B} \in \mathbb{R}^{d\times k}$ with orthonormal columns and $\theta \in \mathbb{R}^d$. Put

$$\widehat{P} = \widehat{B}\widehat{B}^\top, \qquad A(\theta, \widehat{B}) = \|(I_d - \widehat{P})\theta\|_2^2. \tag{13.40}$$

Observe n independent pairs with

$$Y_j = X_j^\top\theta + \varepsilon_j, \qquad X_j \sim \mathcal{N}(0, I_d), \qquad j \in [n]. \tag{13.41}$$

The noises are mutually independent and independent of all covariates, have mean zero, and have common variance σ^2. Let Z have rows $(\widehat{B}^\top X_j)^\top$ and let $Y = (Y_1, \dots, Y_n)^\top$. Define

$$\widehat{a} = (Z^\top Z)^{-1} Z^\top Y, \qquad \widehat{\theta} = \widehat{B}\widehat{a}. \tag{13.42}$$

Theorem 13.A.1 (Exact Gaussian transfer risk)**.** *Under this model, if $n > k + 1$, then*

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] = A(\theta, \widehat{B}) + (\sigma^2 + A(\theta, \widehat{B}))\frac{k}{n - k - 1}. \tag{13.43}$$

For an independent test covariate $X \sim \mathcal{N}(0, I_d)$, the expected prediction error is

$$\mathbb{E}[(X^\top \widehat{\theta} - X^\top \theta)^2] = \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2]. \tag{13.44}$$

Proof. Write $b_\star = \widehat{B}^\top \theta$ and $v = (I_d - \widehat{P})\theta$. Then $\theta = \widehat{B} b_\star + v$ with $\widehat{B}^\top v = 0$, and

$$Y = Z b_\star + \xi, \qquad \xi_j = X_j^\top v + \varepsilon_j.$$

For each j,

$$\mathrm{Cov}(\widehat{B}^\top X_j, X_j^\top v) = \widehat{B}^\top \mathbb{E}[X_j X_j^\top] v = \widehat{B}^\top v = 0.$$

The two projections are jointly Gaussian, so zero covariance makes them independent. The noise ε_j is also independent of X_j. Therefore ξ is independent of Z, has mean zero, and has covariance $(\sigma^2 + \|v\|_2^2) I_n$. Conditional on Z,

$$\mathbb{E}[\|\widehat{a} - b_\star\|_2^2 \mid Z] = (\sigma^2 + \|v\|_2^2)\,\mathrm{tr}((Z^\top Z)^{-1}).$$

The Wishart inverse-mean identity gives, for $n > k + 1$,

$$\mathbb{E}[(Z^\top Z)^{-1}] = \frac{I_k}{n - k - 1}. \tag{13.45}$$

Since $\widehat{B}(\widehat{a} - b_\star)$ is orthogonal to v,

$$\|\widehat{\theta} - \theta\|_2^2 = \|\widehat{a} - b_\star\|_2^2 + \|v\|_2^2.$$

Take expectations. For the independent test covariate X, independence from $\widehat{\theta}$ and $\mathbb{E}[XX^\top] = I_d$ give

$$\mathbb{E}[(X^\top \widehat{\theta} - X^\top \theta)^2 \mid \widehat{\theta}] = (\widehat{\theta} - \theta)^\top \mathbb{E}[XX^\top](\widehat{\theta} - \theta) = \|\widehat{\theta} - \theta\|_2^2.$$

□

If $\widehat{B}$ is random but independent of the target sample, apply the theorem conditionally on $\widehat{B}$ and then average. In the exact-sharing model $\theta = B a_0$, the resulting identity and Eq. (13.35) give Eq. (13.39).

For a known correct representation, $A = 0$ and the risk is $\sigma^2 k/(n - k - 1)$. Ambient ordinary least squares has risk $\sigma^2 d/(n - d - 1)$ when $n > d + 1$. For misspecified sharing $\theta = Ba + r$ with $B^\top r = 0$ and $\widehat{B} = B$, the representation error is $\|r\|_2^2$. Exercise 9 asks you to derive the exact comparison: the restricted estimator has smaller expected error if and only if

$$\|r\|_2^2 < \frac{\sigma^2 (d - k)}{n - d - 1}.$$

The restriction can save variance with little target data, but a fixed nonzero residual eventually outweighs that saving.

13.B Optional comparison: intrinsic input dimension

A single regression problem can become easier because of the input distribution's geometry, without any sharing across tasks. For a pair (X, Y) with $\mathbb{E}[Y^2] < \infty$, let P_X be the input law on $\mathbb{R}^d$ and choose a measurable version m of the regression function, so $m(X) = \mathbb{E}[Y \mid X]$ almost surely. If P_X concentrates near a d^*-dimensional set, a regression rate may depend on d^* rather than d.

A theorem of Kohler, Molinero Römer, and Krzyżak gives such a rate for a specified deep sigmoid estimator: prescribed random initialization, random pruning before training, output truncation, and full-batch gradient descent with data-dependent Armijo-style stepsizes. Include the training sample and the estimator's prescribed initialization and pruning in the same joint experiment. Under their smoothness, tail, architecture, and tuning assumptions, and support in a shrinking neighborhood of a compact Lipschitz manifold, their bound has the form

$$\mathbb{E}[\|\widehat{m}_n - m\|^2_{L^2(P_X)}] \le C'(\log n)^{4d^*L+d+3} n^{-2p/(2p+d^*)}.$$

Here p is smoothness, L is the prescribed depth parameter, and the the tuning uses p, d^*. The tube radius is $n^{-1/(2p+d^*)}(\log n)^{-4L-1}$, not fixed positive thickness. The sufficient widths and iteration counts have large polynomial exponents. This is a theorem for that procedure under those assumptions, not evidence for the manifold hypothesis or a guarantee for ordinary network training.

Classical local methods also obtain intrinsic-dimension rates: Kpotufe's nearest-neighbor result adapts to local dimension, and Kpotufe and Garg's kernel result adapts to local dimension and local Hölder smoothness. Farahmand, Szepesvári, and Audibert developed a related manifold-adaptive program. Their assumptions and smoothness ranges should be compared explicitly, rather than treating all rates as one identical theorem. The neural result's distinctive content is its guarantee for a specified gradient-descent estimator. Low-dimensional input geometry and shared representations across tasks are different statistical assumptions.

References

Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019. https://openaccess.thecvf.com/content_CVPR_2019/html/Aljundi_Task-Free_Continual_Learning_CVPR_2019_paper.html.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000. https://doi.org/10.1613/jair.731.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. https://doi.org/10.1162/089976603321780317.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. https://doi.org/10.1109/TPAMI.2013.50.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. 2021. https://arxiv.org/abs/2104.13478.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. https://doi.org/10.1137/0707001.

Amir-massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 265–272, 2007. https://doi.org/10.1145/1273496.1273530.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. https://www.deeplearningbook.org/.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, 2001. https://doi.org/10.1002/0471221317.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016. https://doi.org/10.1098/rsta.2015.0202.

Michael Kohler, Vincent Molinero Römer, and Adam Krzyżak. Learning of deep neural network regression estimates using gradient descent with pruning. 2026. https://arxiv.org/abs/2608.25743.

Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, 2011. https://proceedings.neurips.cc/paper/2011/hash/05f971b5ec196b8c65b75d2ef8267331-Abstract.html.

Samory Kpotufe and Vikas K. Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems 26*, 2013. https://proceedings.neurips.cc/paper/2013/hash/28fc2782ea7ef51c1104ccf7b9bea13d-Abstract.html.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. https://doi.org/10.1109/ACCESS.2020.3031549.

Timothée Lesort, Natalia Díaz-Rodríguez, Jean-François Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018. https://doi.org/10.1016/j.neunet.2018.07.006.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016. https://www.jmlr.org/papers/v17/15-242.html.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. https://doi.org/10.1109/JPROC.2021.3058954.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems 33*, 2020. https://proceedings.neurips.cc/paper/2020/hash/59587bffec1c7846f3e34230141556ae-Abstract.html.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10434–10443, 2021. https://proceedings.mlr.press/v139/tripuraneni21a.html.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007. https://doi.org/10.1007/s11222-007-9033-z.