

Asymptotically Optimal Regularization in Smooth Parametric Models

Percy Liang
University of California, Berkeley
Berkeley, CA 94720
pliang@cs.berkeley.edu

Francis Bach
INRIA - École Normale Supérieure, France
75214 Paris, France
francis.bach@ens.fr

Guillaume Bouchard
Xerox Research Centre Europe, France
38240 Meylan, France
Guillaume.Bouchard@xrce.xerox.com

Michael I. Jordan
University of California, Berkeley
Berkeley, CA 94720
jordan@cs.berkeley.edu

June 3, 2010

Abstract

Many regularization schemes have been employed in statistical learning, where each is motivated by some assumption about the problem domain. In this paper, we focus on regularizers in smooth parametric models and present an asymptotic analysis that allows us to see how the validity of these assumptions affects the risk of a particular regularized estimator. In addition, our analysis motivates an algorithm for optimizing regularization parameters, which in turn can be analyzed within our framework. We apply our analysis to several examples, including hybrid generative-discriminative learning and multi-task learning.

1 Introduction

Many problems in machine learning and statistics involve the estimation of parameters from finite data. Although empirical risk minimization has favorable limiting properties, it is well known that empirical risk minimization can overfit on finite data. Hence, various forms of regularization (a canonical example being a penalty on the norm of the parameters) have been employed to control this overfitting.

Regularizers are usually chosen based on assumptions about the problem domain at hand. For example, in classification, we might use quadratic regularization if we expect the data to be separable with a large margin. We might regularize with a generative model if we think the generative model is roughly well-specified [10, 30, 23, 25]. In multi-task learning, we might penalize deviation between parameters across tasks if we believe the tasks are similar [3, 17, 2, 19].

In each of these scenarios, we would like (1) a procedure for choosing the parameters of the regularizer (for example, its strength) and (2) an analysis that shows the amount by which regularization reduces expected risk, expressed as a function of the compatibility between the regularizer and the problem domain. In this paper, we address these two points by developing an asymptotic analysis of smooth regularizers for parametric problems. The key idea is to derive a second-order Taylor approximation of the expected risk, yielding a simple and interpretable quadratic form which can be directly minimized with respect to the regularization parameters. We first develop the general theory (Section 2) and then apply it to some examples of common regularizers used in practice (Section 3).

2 General theory

2.1 Notation

We use uppercase letters (e.g., L, R, Z) to denote random variables and script letters (e.g., $\mathcal{L}, \mathcal{R}, \mathcal{I}$) to denote constant limits of random variables.

For a λ -parametrized differentiable function $\theta \mapsto f(\lambda; \theta)$, let \dot{f} , \ddot{f} , and $\overset{\cdot\cdot}{f}$ denote the first, second and third derivatives of f with respect to θ , and let $\nabla f(\lambda; \theta)$ denote the derivative with respect to λ . For a vector v , let $v^\otimes = vv^\top$, and $v^{\otimes 3}$ be the rank-3 tensor formed.

Let v be a vector, A and B be symmetric matrices, and T and U be symmetric rank-3 tensors. Then:

- $T[v]$ is the matrix with entries $T[v]_{ij} = \sum_k T_{ijk}v_k$.
- $T[A]$ is the vector with components $T[A]_i = \sum_{j,k} T_{ijk}A_{jk}$.
- $T[U] = \sum_{i,j,k} T_{ijk}U_{ijk}$.
- $A[B] = \text{tr}\{AB\}$.

The Dirac-delta function is expressed as $\delta_x(y) = 1$ if $x = y$, 0 otherwise.

Let $X_n = O_p(1)$ denote a sequence of random variables which is bounded in probability, that is, for every $\epsilon > 0$, there exists $M < \infty$ such that $\sup_n P(X_n > M) \leq \epsilon$. Let $X_n \xrightarrow{P} X$ denote convergence in probability, that is, for every $\epsilon > 0$, $\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0$. We write $X_n = O_p(Y_n)$ to mean $\frac{X_n}{Y_n} = O_p(1)$. Expectation and variance operators are denoted as $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$, respectively.

2.2 Setup

We are given a *loss function* $\ell(\cdot; \theta)$ parametrized by $\theta \in \mathbb{R}^d$ (e.g., for least squares linear regression, $\ell((x, y); \theta) = \frac{1}{2}(y - x^\top \theta)^2$). Our goal is to minimize the *expected risk* $\mathcal{L}(\theta)$, defined as follows:

$$\theta_\infty \stackrel{\text{def}}{=} \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \mathcal{L}(\theta), \quad \text{where } \mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{Z \sim p^*} [\ell(Z; \theta)], \quad (1)$$

which averages the loss over some true data generating distribution $p^*(Z)$. We do not have access to p^* , but instead receive a sample of n i.i.d. data points Z_1, \dots, Z_n drawn from p^* . The standard *unregularized estimator* minimizes the *empirical risk*:

$$\hat{\theta}_n^0 \stackrel{\text{def}}{=} \underset{\theta \in \mathbb{R}^d}{\text{argmin}} L_n(\theta), \quad \text{where } L_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \ell(Z_i, \theta). \quad (2)$$

Although $\hat{\theta}_n^0$ is consistent (that is, it converges in probability to θ_∞) under relatively weak conditions, it is well known that regularization can improve performance substantially for finite n .

Let $R_n(\lambda, \theta)$ be a (possibly data-dependent) regularization function, where $\lambda \in \mathbb{R}^b$ are the regularization parameters. For linear regression, we might regularize the squared norm of the parameter vector by setting $R_n(\lambda, \theta) = \frac{\lambda}{2n} \|\theta\|^2$, where $\lambda \in \mathbb{R}$ determines the amount of regularization. Note that in general, (i) the regularizer is a random function which can depend on the training data, and (ii) there can be more than one regularization parameter (λ is a vector).

Define the *regularized estimator* as follows:

$$\hat{\theta}_n^\lambda \stackrel{\text{def}}{=} \underset{\theta \in \mathbb{R}^d}{\text{argmin}} L_n(\theta) + R_n(\lambda, \theta). \quad (3)$$

Assume $R_n(0, \theta) \equiv 0$ ($\lambda = 0$ corresponds to no regularization) and $R_n(\lambda, \theta) \xrightarrow{P} 0$ as $n \rightarrow \infty$ (the regularization vanishes with the amount of training data tending to infinity).

The goal of this paper is to choose good values of λ and analyze the subsequent impact on performance. Specifically, we wish to minimize the *relative risk*:

$$\mathbb{L}_n(\lambda) \stackrel{\text{def}}{=} \mathbb{E}_{Z_1, \dots, Z_n \sim p^*} [\mathcal{L}(\hat{\theta}_n^\lambda) - \mathcal{L}(\hat{\theta}_n^0)], \quad (4)$$

which is the difference in risk (averaged over the training data) between the regularized and unregularized estimators; $\mathbb{L}_n(\lambda) < 0$ is desirable. Clearly, $\text{argmin}_\lambda \mathbb{L}_n(\lambda)$ is the asymptotically optimal data-independent regularization parameter. However, it is difficult to get a handle on $\mathbb{L}_n(\lambda)$, let alone optimize it. Therefore, the main focus of this work is on deriving an asymptotic expansion for $\mathbb{L}_n(\lambda)$ that is tractable. This will enable us to analyze and compare different regularizers.

In this paper, we make the following assumptions:

Assumption 1 (Compact support). *The true distribution $p^*(Z)$ has compact support.*

Assumption 2 (Smooth loss). *The loss function $\ell(z, \theta)$ is thrice-differentiable with respect to θ . Furthermore, assume the expected Hessian of the loss function is positive definite ($\check{\mathcal{L}}(\theta_\infty) \succ 0$).¹*

Assumption 3 (Smooth regularizer). *The regularizer $R_n(\lambda, \theta)$ is thrice-differentiable with respect to θ and differentiable with respect to λ .*

While we do not explicitly assume convexity of ℓ and R_n , the local nature of our subsequent asymptotic analysis will mean that we are essentially working in a strongly convex setting.

In general, we assume nothing about the relationship between the data generating distribution p^* and the loss function ℓ . At certain points in this paper, we will be able to obtain stronger results by considering a specialization of our results to the setting where the model is well-specified:

Definition 1 (Well-specified model). *The loss function ℓ (model) is well-specified with respect to the data generating distribution $p^*(z)$ if we can write $\ell(z; \theta) = -\log p(z_2 | z_1; \theta)$ for some decomposition $z = (z_1, z_2)$ such that $p^*(z_2 | z_1) \equiv p_{\theta_\infty}(z_2 | z_1)$.*

The typical setting is when z_1 is the input and z_2 is the output, in which case we call $p(z_2 | z_1; \theta)$ a *discriminative model*. We also allow z_1 to be degenerate, in which case we call $p(z_2 | z_1; \theta)$ a *generative model*.

2.3 Rate of regularization strength

Before tackling the question of how to choose the best regularization parameter λ , let us establish some basic properties that the regularizer $R_n(\lambda, \theta)$ ought to satisfy. First, a desirable property is consistency ($\hat{\theta}_n^\lambda \xrightarrow{P} \theta_\infty$), i.e., convergence of the estimated parameters to the parameters that achieve the minimum possible risk in our hypothesis class. To achieve this, it suffices (and in general also necessitates) that (1) the loss class satisfies standard uniform convergence properties [33] and (2) the regularizer has vanishing strength in the limit of infinite data ($R_n(\lambda, \theta) \xrightarrow{P} 0$). These two properties can be verified given Assumptions 1 and 2.

The next question is at what rate $R_n(\lambda, \theta)$ should converge to 0? The following theorem establishes that $O_p(\frac{1}{n})$ is the optimal rate:

Theorem 1. *In general, $R_n(\lambda, \theta) = O_p(n^{-1})$ is the rate that minimizes the asymptotic relative risk $\mathbb{L}(\lambda)$.*

See Appendix A for the proof. Here is the rough sketch: Suppose that $R_n = O_p(a_n)$ for some constant sequence a_n . In Appendix A, we will derive following expansion $\mathbb{L}_n(\lambda) = O_p(n^{-1})O_p(a_n) + O_p(1)O_p(a_n^2) + O_p(a_n^3)$. In order to minimize $\mathbb{L}_n(\lambda)$, we need the first two terms to be on the same order (since the two will typically be of opposite signs), which can be achieved with $a_n = n^{-1}$. Of course if the regularizer is degenerate, e.g., $R_n(\theta) = a_n(1 - \delta_{\theta_\infty}(\theta))$, then letting $a_n = \infty$ is optimal; we ignore these degenerate cases.

¹This assumption can be weakened. If $\check{\mathcal{L}} \neq 0$, the parameters can only be estimated up to the row space of $\check{\mathcal{L}}$. But since we are interested in the parameters θ only in terms of $\mathcal{L}(\theta)$, this type of non-identifiability of the parameters is irrelevant.

With this rate, it is natural to regard the regularizer as a prior $p(\theta \mid \lambda) \propto \exp\{-R_n(\lambda, \theta)\}$ if R_n is non-random (does not depend on the training data) with λ serving as the hyperparameters of the model. If the loss is the negative log-likelihood of some (not necessarily well-specified) model ($\ell(z, \theta) = -\log p(z_2 \mid z_1; \theta)$), then $\hat{\theta}_n^\lambda$ is the maximum a posteriori (MAP) estimate. In this light, our work can be viewed as a method for setting priors in order to optimize predictive performance. It would be interesting to compare our method with methods for setting priors based on objective Bayesian principles, for example, reference priors [7].

2.4 Asymptotic expansion of the relative risk

Basic definitions:		
$\theta \in \mathbb{R}^d$		parameter vector
$z \in \mathcal{Z}$		data point (e.g., $z = (x, y)$ for prediction problems)
$p^*(z)$		true data generating distribution
$\ell(z; \theta)$		loss function
$Z_1, \dots, Z_n \sim p^*$		training data points
$L_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \ell(Z_i; \theta)$		empirical risk
$R_n(\lambda; \theta)$		regularizer
$\mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathbb{E}[\ell(Z; \theta)]$		empirical risk
$\hat{\theta}_n^0 = \operatorname{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta)$		unregularized estimator
$\hat{\theta}_n^\lambda = \operatorname{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta) + R_n(\lambda, \theta)$		regularized estimator with regularization parameter λ
$\theta_\infty = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$		optimal parameters (also, $\hat{\theta}_n^\lambda \xrightarrow{P} \theta_\infty$)
Population limits (note all evaluations are at θ_∞):		
\mathcal{L}	$\stackrel{\text{def}}{=} \mathbb{E}[\ell(Z; \theta_\infty)] \in \mathbb{R}$	Bayes risk
$\dot{\mathcal{L}}$	$\stackrel{\text{def}}{=} \mathbb{E}[\dot{\ell}(Z; \theta_\infty)] = 0 \in \mathbb{R}^d$	Expected gradient of loss (score function)
$\ddot{\mathcal{L}}$	$\stackrel{\text{def}}{=} \mathbb{E}[\ddot{\ell}(Z; \theta_\infty)] \in \mathbb{R}^{d \times d}$	Expected Hessian of loss
$\ddot{\mathcal{L}}$	$\stackrel{\text{def}}{=} \mathbb{E}[\ddot{\ell}(Z; \theta_\infty)] \in \mathbb{R}^{d \times d \times d}$	Expected third-derivative of loss
$\mathcal{I}_{\ell\ell}$	$\stackrel{\text{def}}{=} \mathbb{E}[\dot{\ell}(Z; \theta_\infty) \otimes \dot{\ell}(Z; \theta_\infty)] \in \mathbb{R}^{d \times d}$	Fisher information (variance of score)
$\mathcal{I}_{\ell^2\ell}(\lambda)$	$\stackrel{\text{def}}{=} \mathbb{E}[\ddot{\ell}(Z; \theta_\infty) \otimes \dot{\ell}(\theta_\infty)] \in \mathbb{R}^{d \times d \times d}$	Expected loss Hessian-gradient product
$\mathcal{R}(\lambda)$	$\stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n \cdot \mathbb{E}[R_n(\lambda, \theta_\infty)] \in \mathbb{R}$	Expected regularizer
$\dot{\mathcal{R}}(\lambda)$	$\stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n \cdot \mathbb{E}[\dot{R}_n(\lambda, \theta_\infty)] \in \mathbb{R}^d$	Expected gradient of regularizer
$\ddot{\mathcal{R}}(\lambda)$	$\stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n \cdot \mathbb{E}[\ddot{R}_n(\lambda, \theta_\infty)] \in \mathbb{R}^{d \times d}$	Expected Hessian of regularizer
$\mathcal{I}_{\ell r}(\lambda)$	$\stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n^2 \cdot \mathbb{E}[\dot{L}_n(\theta_\infty) \dot{R}_n(\lambda, \theta_\infty)^\top] \in \mathbb{R}^{d \times d}$	Random loss-regularizer alignment
\mathcal{B}	$\stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n \cdot \mathbb{E}[\hat{\theta}_n^0 - \theta_\infty] \in \mathbb{R}^d$	Asymptotic bias of unregularized estimator

Table 1: Expectations are taken with respect to the training set, with $Z \sim p^*$ denoting a generic sample from the true distribution. Derivatives are taken with respect to the parameter θ evaluated at the limiting parameters θ_∞ (around which we perform the asymptotic expansion). We will start to omit θ_∞ (e.g., writing \dot{L}_n for $\dot{L}_n(\theta_\infty)$ and $\ddot{R}_n(\lambda)$ for $\ddot{R}_n(\lambda, \theta_\infty)$) when there is no confusion.

Our main result is the following theorem, which provides a simple interpretable asymptotic expression for the relative risk (see Appendix A for the proof):

Theorem 2. *Assume $R_n(\lambda, \theta_\infty) = O_p(n^{-1})$. The relative risk admits the following asymptotic expansion:*

$$\mathbb{L}_n(\lambda) = \mathbb{L}(\lambda) \cdot n^{-2} + O_p(n^{-\frac{5}{2}}), \quad (5)$$

where $\mathbb{L}(\lambda)$ is the asymptotic relative risk:

$$\mathbb{L}(\lambda) \stackrel{\text{def}}{=} \frac{1}{2} \operatorname{tr}\{\dot{\mathcal{R}}(\lambda) \otimes \ddot{\mathcal{L}}^{-1}\} - \operatorname{tr}\{\mathcal{I}_{\ell\ell} \ddot{\mathcal{L}}^{-1} \ddot{\mathcal{R}}(\lambda) \ddot{\mathcal{L}}^{-1}\} - 2\mathcal{B}^\top \dot{\mathcal{R}}(\lambda) + \operatorname{tr}\{\mathcal{I}_{\ell r}(\lambda) \ddot{\mathcal{L}}^{-1}\}, \quad (6)$$

where the various quantities are defined in Table 1.

The significance of Theorem 2 is writing the relative risk $\mathbb{L}_n(\lambda)$ (the quantity we want to minimize but difficult to analyze directly) in terms of the asymptotic relative risk $\mathbb{L}(\lambda)$ (6), which contains terms which we will be able to interpret naturally and optimize. We emphasize that (6) is the most important equation in this paper.

Now we proceed to interpret the asymptotic relative risk $\mathbb{L}(\lambda)$. First, note that the asymptotic relative risk operates at the $O(n^{-2})$ scale. This is natural, because maximum likelihood estimators have expected excess risk $O(n^{-1})$ and are asymptotically optimal up to this first order; our asymptotic improvements therefore come at the next order.

The asymptotic relative risk $\mathbb{L}(\lambda)$ consists of three parts. We will interpret each one in turn:

- *Squared bias* of the regularizer $\text{tr}\{\dot{\mathcal{R}}(\lambda)^\otimes \ddot{\mathcal{L}}^{-1}\}$: $\dot{\mathcal{R}}(\lambda)$ is the gradient of the regularizer at the limiting parameters θ_∞ . If we were sitting at θ_∞ , this would be the direction that the regularizer would push us away. Therefore, regularizers with small values are desirable.

$\ddot{\mathcal{L}}$ is the expected Hessian of the loss function at θ_∞ , which intuitively defines an Mahalanobis metric on the parameter space. Directions of large magnitude according to $\ddot{\mathcal{L}}$ (e.g., $v \in \mathbb{R}^d$ such that $\|\ddot{\mathcal{L}}v\|$ is large) are easier to estimate.

The squared bias of the regularizer, which can also be written as $\dot{\mathcal{R}}(\lambda)^\top \ddot{\mathcal{L}}^{-1} \dot{\mathcal{R}}(\lambda)$, is the squared norm of $\dot{\mathcal{R}}(\lambda)$ with respect to the Mahalanobis metric given by $\ddot{\mathcal{L}}$. Note that the squared regularizer bias is always positive; thus it always increases the risk by an amount which depends on how “wrong” the regularizer is.

- *Variance reduction* provided by the regularizer $\text{tr}\{\mathcal{I}_{\ell\ell} \ddot{\mathcal{L}}^{-1} \ddot{\mathcal{R}}(\lambda) \ddot{\mathcal{L}}^{-1}\}$: The key quantity is $\ddot{\mathcal{R}}(\lambda)$, the Hessian of the regularizer, which is the amount of stability provided by the regularizer (larger is better). For convex regularizers, $\ddot{\mathcal{R}}(\lambda) \succeq 0$.

The impact of $\ddot{\mathcal{R}}(\lambda)$ is channeled through $\ddot{\mathcal{L}}^{-1}$ and $\mathcal{I}_{\ell\ell}$. If the model is *well-specified*, then $\mathcal{I}_{\ell\ell} = \ddot{\mathcal{L}}$ by the first Bartlett identity [4] (see Proposition 2), and the variance reduction term simplifies to $\text{tr}\{\ddot{\mathcal{R}}(\lambda) \ddot{\mathcal{L}}^{-1}\}$.

- *Alignment between regularizer bias and unregularized estimator bias* $2\mathcal{B}^\top \dot{\mathcal{R}}(\lambda) - \text{tr}\{\mathcal{I}_{\ell r}(\lambda) \ddot{\mathcal{L}}^{-1}\}$: The alignment has two parts.

The first part consists of the dot product between \mathcal{B} , the asymptotic bias (the expected parameter error $\hat{\theta}_n^0 - \theta_\infty$ of the unregularized estimator scaled up by n), and $\dot{\mathcal{R}}(\lambda)$, the gradient of the regularizer. A positive dot product means that the direction the regularizer is pushing away from ($\dot{\mathcal{R}}(\lambda)$) is aligned with the direction that the unregularized estimator consistently errs in (\mathcal{B}), which is good. Note that in some cases, such as linear regression, the unregularized estimator (least squares) is unbiased ($\mathcal{B} = 0$).

The second part consists of $\mathcal{I}_{\ell r} \ddot{\mathcal{L}}^{-1}$, which measures the dot product (in the space defined by the metric $\ddot{\mathcal{L}}$) between the direction of increasing loss (\dot{L}_n) and the direction the regularizer is pushing away from (\dot{R}_n). A negative dot product means that the regularizer compensates for the loss, which is good. Note that $\mathcal{I}_{\ell r} = 0$ when the regularizer R_n is non-random (i.e., does not depend on the training data).

In summary, the main equation (6) is largely about an asymptotic bias-variance tradeoff, as governed by the first two terms of (6). However, second-order asymptotics generates many many dependencies via cross terms of Taylor expansions, which are responsible for the additional alignment terms.

Theorem 2 provides us with an expression for the asymptotic relative risk in terms of various limiting quantities defined in Table 1 (e.g., $\ddot{\mathcal{L}}$, $\ddot{\mathcal{R}}$, \mathcal{B} , etc.). When we specialize to examples in Section 3, we will need to compute these quantities in terms of the loss function ℓ and regularizer R_n . The most complicated part of the asymptotic risk in (6) is the asymptotic bias \mathcal{B} of the unregularized estimator. Therefore, we express \mathcal{B} in terms of various derivatives:

Proposition 1 (Asymptotic bias of unregularized estimator). *The asymptotic bias $\mathcal{B} = \mathbb{E}[\hat{\theta}_n^0 - \theta_\infty]$ is*

$$\mathcal{B} = \ddot{\mathcal{L}}^{-1} \mathcal{I}_{\ell^2 \ell} [\ddot{\mathcal{L}}^{-1}] - \frac{1}{2} \ddot{\mathcal{L}}^{-1} \ddot{\mathcal{L}} [\ddot{\mathcal{L}}^{-1} \mathcal{I}_{\ell \ell} \ddot{\mathcal{L}}^{-1}]. \quad (7)$$

Proof: see Appendix B.

2.4.1 Simplifications

If the regularizer has the simplified form $R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta)$ for some function $r(\theta)$, then the expected derivatives involving the regularizer simplify accordingly:

$$\dot{\mathcal{R}} = \lambda \dot{r}, \quad \ddot{\mathcal{R}} = \lambda \ddot{r}, \quad \mathcal{I}_{\ell r} = 0. \quad (8)$$

If the model is well-specified (Definition 1), then we have:

Proposition 2 (First Bartlett identity). *If the loss function ℓ is well-specified with respect to the data generating distribution p^* , then $\mathcal{I}_{\ell \ell} = \ddot{\mathcal{L}}$.*

Proof: see Appendix C. The consequence of Proposition 2 is that the second term of (6) simplifies to $-\text{tr}\{\ddot{\mathcal{R}}(\lambda) \ddot{\mathcal{L}}^{-1}\}$.

2.5 Oracle regularizer

The principal advantage of having a simple expression for the asymptotic relative risk $\mathbb{L}(\lambda)$ given by Theorem 2 is that we can minimize it with respect to λ . In particular, let

$$\lambda^* \stackrel{\text{def}}{=} \underset{\lambda}{\text{argmin}} \mathbb{L}(\lambda) \quad (9)$$

be the asymptotically optimal data-independent regularization parameter, which we call the *oracle regularization parameter*. Let $\hat{\theta}_n^{\lambda^*}$ be the corresponding *oracle estimator* that uses that optimal value λ^* .

We have a closed form for λ^* in the important special case that the regularization parameter λ is a scalar denoting the strength of the regularizer:

Corollary 1 (Oracle regularization strength). *If $R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta)$ for some $r(\theta)$, then*

$$\lambda^* = \underset{\lambda}{\text{argmin}} \mathbb{L}(\lambda) = \frac{\text{tr}\{\mathcal{I}_{\ell \ell} \ddot{\mathcal{L}}^{-1} \ddot{r} \ddot{\mathcal{L}}^{-1}\} + 2\mathcal{B}^\top \dot{r}}{\dot{r}^\top \ddot{\mathcal{L}}^{-1} \dot{r}} \stackrel{\text{def}}{=} \frac{\mathcal{C}_1}{\mathcal{C}_2}, \quad \mathbb{L}(\lambda^*) = -\frac{\mathcal{C}_1^2}{2\mathcal{C}_2}. \quad (10)$$

For simplicity, we have assumed a non-random regularizer of the form $R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta)$. We use \mathcal{C}_1 and \mathcal{C}_2 to denote the numerator (variance plus alignment) and the denominator (squared bias), respectively.

Proof. Note that (6) is a quadratic function of λ . Differentiate with respect to λ , set to zero and solve. Compute $\mathbb{L}(\lambda^*)$ by substituting λ^* from (10) back into (6). \square

In general, λ^* will depend on θ_∞ ; hence it is not computable from data. Section 2.6 will remedy this. Nevertheless, the oracle regularizer provides an upper bound on asymptotic performance and sheds some insight into the relevant quantities that make a regularizer useful.

Note $\mathbb{L}(\lambda^*) \leq 0$, since optimizing λ^* must be no worse than not regularizing since $\mathbb{L}(0) = 0$. What might be surprising at first is that the oracle regularization strength λ^* can be negative (corresponding to “anti-regularization”). If $\frac{\partial \mathbb{L}(\lambda)}{\partial \lambda} = -\mathcal{C}_1 < 0$, however, then some amount of (positive) regularization is guaranteed to help. In this case, $\lambda^* > 0$, and $\mathbb{L}(\lambda) < 0$ for $0 < \lambda < 2\lambda^*$.

2.6 Plugin regularizer

While the oracle regularizer $R_n(\lambda^*, \theta)$ given by (10) is asymptotically optimal, λ^* depends on the unknown θ_∞ , so $\hat{\theta}_n^{\lambda^*}$ is actually correspond to an estimator that is *not* implementable. In this section, we develop the plugin regularizer as a way to avoid the dependence on θ_∞ . The key idea is to substitute λ^* with a noisy estimate $\hat{\lambda}_n$. We define the *plugin regularization parameter*:

$$\hat{\lambda}_n \stackrel{\text{def}}{=} \lambda^* + \varepsilon_n, \quad (11)$$

where $\varepsilon_n = O_p(n^{-\frac{1}{2}})$ is some noise. We will describe how to obtain a plugin regularization parameter shortly, but first let us see the consequences of having one: Define the *plugin estimator* as follows:

$$\hat{\theta}_n^{\hat{\lambda}_n} \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} L_n(\theta) + R_n(\hat{\lambda}_n, \theta). \quad (12)$$

How well does this plugin estimator work—that is, what is its relative risk $\mathbb{E}[\mathcal{L}(\hat{\theta}_n^{\hat{\lambda}_n}) - \mathcal{L}(\hat{\theta}_n^0)]$? Unfortunately, we cannot simply write $\mathbb{L}_n(\hat{\lambda}_n)$ and apply Theorem 2 because the relative risk function $\mathbb{L}(\cdot)$ can only be applied to non-random regularization parameters; in contrast, $\hat{\lambda}_n$ depends on the data.

A new regularizer However, we can still leverage existing machinery by defining a new *plugin regularizer*:

$$R_n^\bullet(\lambda^\bullet, \theta) \stackrel{\text{def}}{=} \lambda^\bullet R_n(\hat{\lambda}_n, \theta) \quad (13)$$

with regularization parameter $\lambda^\bullet \in \mathbb{R}$, which is defined in terms of the original regularizer R_n . Note that R_n^\bullet depends on the data through $\hat{\lambda}_n$, and this dependence is one of the main reasons we needed to support the generality of a random regularizer. Henceforth, the superscript \bullet will denote quantities defined in terms of the plugin regularizer R_n^\bullet rather than the original regularizer R_n .

The estimator that uses the plugin regularizer with regularization parameter λ^\bullet is defined as follows:

$$\hat{\theta}_n^{\lambda^\bullet} \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} L_n(\theta) + R_n^\bullet(\lambda^\bullet, \theta) \quad (14)$$

The purpose of introducing the plugin regularizer R_n^\bullet and its associated estimator $\hat{\theta}_n^{\lambda^\bullet}$ is the following key identity:

$$\hat{\theta}_n^{\hat{\lambda}_n} = \hat{\theta}_n^{\lambda^\bullet=1}, \quad (15)$$

where we relate the original plugin estimator we wished to study on the left-hand side to an estimator that has a non-random regularization parameter (namely, $\lambda^\bullet = 1$) on the right-hand side. Notably, we know how to compute the relative risk for estimators involving non-random regularization parameters.

Define the relative risk of estimators using the plugin regularizer (note the analogy with (4)):

$$\mathbb{L}_n^\bullet(\lambda^\bullet) \stackrel{\text{def}}{=} \mathbb{E}[\mathcal{L}(\hat{\theta}_n^{\lambda^\bullet}) - \mathcal{L}(\hat{\theta}_n^0)]. \quad (16)$$

In particular, we are interested in $\lambda^\bullet = 1$, since the relative risk of $\mathbb{L}_n^\bullet(1)$ is exactly the relative risk of the plugin estimator $\hat{\theta}_n^{\hat{\lambda}_n}$ by (15).

Further improvements Instead of settling for $\lambda^\bullet = 1$, we could try to squeeze more out of the plugin regularizer by further optimizing λ^\bullet as follows:

$$\lambda^{\bullet*} \stackrel{\text{def}}{=} \underset{\lambda^\bullet}{\operatorname{argmin}} \mathbb{L}_n^\bullet(\lambda^\bullet), \quad (17)$$

which leads to the *oracle plugin estimator* $\hat{\theta}_n^{\lambda^{\bullet*}}$.

In general, this is not useful since $\lambda^{\bullet*}$ might depend on θ_∞ , and the whole point of plugin is to remove this dependence. However, in a fortuitous turn of events, for some models as we will see in Sections 3.1 and 3.5, $\lambda^{\bullet*}$ is in fact independent of θ_∞ . When this is the case, $\hat{\theta}_n^{\lambda^{\bullet*}}$ actually *is* implementable. Table 2 summarizes all the estimators we have discussed.

Having defined all the relevant plugin-related estimators, we turn to computing and comparing their relative risks. The main question is whether the implementable procedure (the plugin estimator $\hat{\theta}_n^{\hat{\lambda}_n} = \hat{\theta}_n^{\bullet 1}$) can work “almost as well” as the unimplementable procedure (the oracle estimator $\hat{\theta}_n^{\lambda^*}$). The following theorem answers exactly that:

Theorem 3 (Relative risk of plugin). *The relative risk of the plugin estimator is*

$$\mathbb{L}^\bullet(1) = \mathbb{L}(\lambda^*) + \mathcal{E}, \quad (18)$$

where

$$\mathcal{E} \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n^2 \mathbb{E}[\text{tr}\{\dot{L}_n(\nabla \dot{R}_n(\lambda^*) \varepsilon_n)^\top \ddot{L}^{-1}\}], \quad (19)$$

where $\nabla \dot{R}_n(\lambda^*) \in \mathbb{R}^{d \times b}$. Furthermore, if $R_n(\lambda)$ is linear in λ , then the relative risk of the oracle plugin estimator is

$$\mathbb{L}^\bullet(\lambda^{\bullet*}) = \mathbb{L}^\bullet(1) + \frac{\mathcal{E}^2}{4\mathbb{L}(\lambda^*)}, \quad (20)$$

with the oracle plugin regularization parameter

$$\lambda^{\bullet*} = 1 + \frac{\mathcal{E}}{2\mathbb{L}(\lambda^*)}. \quad (21)$$

Proof: see Appendix D.

Note that the sign of \mathcal{E} depends on the nature of the error ε_n , so PLUGIN could be either better or worse than ORACLE. On the other hand, ORACLEPLUGIN is always better than PLUGIN.

Obtaining a plugin regularization parameter $\hat{\lambda}_n$ So far, we have analyzed the plugin estimator assuming we have an estimate $\hat{\lambda}_n = \lambda^* + \varepsilon_n$. Now, let us consider a useful and general recipe for obtaining $\hat{\lambda}_n$. Note that (10) provides us an expression for λ^* in terms of θ_∞ . Let f denote this function. We can plugin the unregularized estimator $\hat{\theta}_n^0$ (or any other consistent estimator of θ^*) and set $\hat{\lambda}_n = f(\hat{\theta}_n^0)$.

Figure 1 summarizes the resulting algorithm. The algorithm can be viewed as performing adaptive regularization, where a preliminary estimate is used to determine the appropriate amount of regularization, and then a regularized optimization problem is solved to obtain the final estimate.

PLUGIN algorithm:

1. $\hat{\theta}_n^0 = \text{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta)$ [compute unregularized estimate]
2. $\hat{\lambda}_n = f(\hat{\theta}_n^0)$ [compute plugin regularization parameter]
3. $\hat{\theta}_n^{\hat{\lambda}_n} = \text{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta) + R_n(\hat{\lambda}_n, \theta)$ [compute regularized estimate]

Figure 1: Our proposed two-stage algorithm for setting the regularization parameter. First, we use obtain crude preliminary estimate $\hat{\theta}_n^0$ without using regularization. Then we estimate an appropriate regularization parameter $\hat{\lambda}_n$ by using a function f such that $\lambda^* = f(\theta_\infty)$. Finally, we re-solve the optimization problem with regularization to get $\hat{\theta}_n^{\hat{\lambda}_n}$, which is our final answer.

In this case, we can specialize \mathcal{E} (19):

Estimator	Description	Notation	Relative risk
UNREGULARIZED	No regularization	$\hat{\theta}_n^0$	0
ORACLE	Regularization with optimal reg. parameter λ^*	$\hat{\theta}_n^{\lambda^*}$	$\mathbb{L}(\lambda^*)$
PLUGIN	Regularization with plugin reg. parameter $\hat{\lambda}_n$	$\hat{\theta}_n^{\hat{\lambda}_n} = \hat{\theta}_n^{\bullet 1}$	$\mathbb{L}^\bullet(1)$
ORACLEPLUGIN	Optimize strength of the random reg.	$\hat{\theta}_n^{\bullet \lambda^{\bullet \bullet}}$	$\mathbb{L}^\bullet(\lambda^{\bullet \bullet})$

Table 2: Notation for the various estimators and their relative risks.

Theorem 4. Suppose $\lambda^* = f(\theta_\infty)$ for some differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}^b$. If $\hat{\lambda}_n = f(\hat{\theta}_n^0)$, then the results of Theorem 3 hold with

$$\mathcal{E} = -\text{tr}\{\mathcal{I}_{\ell\ell}\ddot{\mathcal{L}}^{-1}\nabla\dot{\mathcal{R}}(\lambda^*)\dot{f}\ddot{\mathcal{L}}^{-1}\}. \quad (22)$$

Proof: see Appendix E.

Table 2 summarizes the estimators that we have defined: the unregularized estimator, UNREGULARIZED ($\hat{\theta}_n^0$); the oracle regularized estimator, ORACLE ($\hat{\theta}_n^{\lambda^*}$); the plugin regularized estimator, PLUGIN ($\hat{\theta}_n^{\hat{\lambda}_n}$; equivalently, $\hat{\theta}_n^{\bullet 1}$); and the oracle plugin regularized estimator, ORACLEPLUGIN ($\hat{\theta}_n^{\bullet \lambda^{\bullet \bullet}}$).

3 Some applications of the theory

In this section, we apply our results from Section 2 to specific problems. Having made all the asymptotic derivations in the general setting, we now only need to make a few straightforward calculations to obtain the asymptotic relative risks and regularization parameters for a given problem. To get some intuition for the theory, we first explore two classical examples from statistics, Gaussian mean estimation (Section 3.1) and binomial estimation (Section 3.2). Then we consider two important examples in machine learning, hybrid generative-discriminative learning (Section 3.4) and multi-task learning (Section 3.5).

3.1 Gaussian mean estimation

Setup A classical problem in statistics is estimating the mean of a distribution from samples. In this work, we assume data are generated from a multivariate Gaussian distribution with d independent components ($p^* = \mathcal{N}(\theta_\infty, I)$), where the mean $\theta_\infty \in \mathbb{R}^d$ is unknown.

We use the negative log-likelihood of the Gaussian distribution as the loss function, $\ell(x; \theta) = \frac{1}{2}(x - \theta)^2$, so we are working in the well-specified setting. In this case, the unregularized (maximum likelihood) estimator is simply the empirical mean:

$$\hat{\theta}_n^0 = \bar{X} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n X_i \quad (23)$$

Consider a regularizer which penalizes the squared norm, heretofore known as the *quadratic regularizer*:

$$R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta), \quad r(\theta) = \frac{1}{2} \|\theta\|^2. \quad (24)$$

Oracle regularization To compute the oracle regularizer (10), we need to first compute the various derivatives of the loss and regularizer: $\dot{L}_n = \theta_\infty - \bar{X}$, $\dot{\mathcal{L}} = I$, $\mathcal{B} = 0$, $\dot{r} = \theta_\infty$, and $\ddot{r} = I$. Note that since the model is well-specified, we also have $\mathcal{I}_{\ell\ell} = \dot{\mathcal{L}}$.

By (10), we can compute the oracle regularization strength and its associated asymptotic relative risk:

$$\lambda^* = \frac{d}{\|\theta_\infty\|^2}, \quad \mathbb{L}(\lambda^*) = -\frac{d^2}{2\|\theta_\infty\|^2}. \quad (25)$$

The form of λ^* is intuitive: the closer θ_∞ is to zero, the more we should regularize towards zero.

Plugin regularization However, λ^* depends on θ_∞ , so let us use the plugin estimator PLUGIN from Section 2.6. By (10), we have

$$f(\theta) = \frac{d}{\|\theta\|^2} \quad \text{and} \quad \dot{f}(\theta) = \frac{-2d\theta}{\|\theta\|^4}, \quad (26)$$

so the plugin estimator, written out explicitly, is defined as follows:

$$\hat{\theta}_n^{\hat{\lambda}_n} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^n (X_i - \theta)^2 + \frac{1}{2} f(\bar{X}) \|\theta\|^2, \quad (27)$$

$$(28)$$

where $f(\bar{X})$ is the estimated regularization strength.

Now let us analyze the relative risk of the plugin estimator $\hat{\theta}_n^{\hat{\lambda}_n} = \hat{\theta}_n^{\bullet 1}$. From Theorem 4, we compute

$$\mathcal{E} = \operatorname{tr} \left\{ \frac{2d\theta_\infty}{\|\theta_\infty\|^4} \theta_\infty^\top \right\} \quad (29)$$

$$= \frac{2d}{\|\theta_\infty\|^2}. \quad (30)$$

Plugging \mathcal{E} into Theorem 3 yields the relative risk for the plugin estimator:

$$\mathbb{L}^\bullet(1) = \mathbb{L}(\lambda^*) + \mathcal{E} \quad (31)$$

$$= -\frac{d^2}{2\|\theta_\infty\|^2} + \frac{2d}{\|\theta_\infty\|^2} \quad (32)$$

$$= -\frac{d(d-4)}{2\|\theta_\infty\|^2}. \quad (33)$$

Note that since $\mathcal{E} > 0$, PLUGIN is always (asymptotically) worse than ORACLE but better than UNREGULARIZED if $d > 4$.

Oracle plugin regularization The implied plugin regularizer is:

$$R_n^\bullet(\theta) = \frac{1}{2} \frac{d\|\theta\|^2}{\|\bar{X}\|^2}. \quad (34)$$

We can optimize over its regularization parameter λ^\bullet to get ORACLEPLUGIN. By Theorem 3, the optimum is attained at

$$\lambda^{\bullet*} = 1 - \frac{\mathcal{E}}{2\mathbb{L}(\lambda^*)} \quad (35)$$

$$= 1 - \frac{2d\|\theta_\infty\|^2}{\|\theta_\infty\|^2 d^2} \quad (36)$$

$$= 1 - \frac{2}{d}. \quad (37)$$

It is important that $\lambda^{\bullet*}$ does not depend on θ_∞ , for otherwise, the oracle plugin estimator would not be implementable. The relative risk, by Theorem 3, is

$$\mathbb{L}^\bullet(\lambda^{\bullet*}) = -\frac{(d-2)^2}{2\|\theta_\infty\|^2}. \quad (38)$$

Note that ORACLEPLUGIN offers a small improvement over PLUGIN. It is superior to UNREGULARIZED when $d > 2$ (by comparison, PLUGIN is superior only when $d > 4$).

The final estimator, written out explicitly, is as follows:

$$\hat{\theta}_n^{\bullet\bullet*} = \operatorname{argmin}_{\theta} L_n(\theta) + \lambda_n^{\bullet\bullet*} R_n^{\bullet\bullet}(\theta) \quad (39)$$

$$= \operatorname{argmin}_{\theta} \frac{1}{2n} \sum_{i=1}^n (X_i - \theta)^2 + \frac{(1 - \frac{2}{d})}{n} \frac{d \|\theta\|^2}{2 \|\bar{X}\|^2}. \quad (40)$$

We can solve this problem in closed form. Differentiating and setting to zero:

$$\hat{\theta}_n^{\bullet\bullet*} - \bar{X} + \frac{d-2}{n \|\bar{X}\|^2} \hat{\theta}_n^{\bullet\bullet*} = 0. \quad (41)$$

Rearranging yields:

$$\hat{\theta}_n^{\bullet\bullet*} = \bar{X} \left(1 - \frac{d-2}{n \|\bar{X}\|^2 + d-2} \right). \quad (42)$$

Comparison with James-Stein In his seminal 1961 paper [20], Stein showed the rather counterintuitive result that the unregularized estimator $\hat{\theta}_n^0$ is not admissible for $d > 2$, despite the fact that the d means being estimated are independent.

Specifically, the James-Stein estimator (JAMESSTEIN) is defined as follows:

$$\hat{\theta}_n^{\text{JS}} \stackrel{\text{def}}{=} \bar{X} \left(1 - \frac{d-2}{n \|\bar{X}\|^2} \right) \quad (43)$$

It was shown in [20] that JAMESSTEIN achieves strictly lower risk than UNREGULARIZED—that is,

$$\mathbb{E}[\mathcal{L}(\hat{\theta}_n^{\text{JS}})] < \mathbb{E}[\mathcal{L}(\hat{\theta}_n^0)] \quad (44)$$

for all n and θ_∞ provided the dimensionality $d > 2$.

If we compare ORACLEPLUGIN in (42) with JAMESSTEIN in (43), we find that the two are essentially the same. Formally, it can be verified that

$$\hat{\theta}_n^{\bullet\lambda\bullet\bullet*} - \hat{\theta}_n^{\text{JS}} = O_p(n^{-\frac{5}{2}}). \quad (45)$$

Both have the intuition we shrink the means towards each other by some factor that reduces variance at expense of only a small increase in bias, which is exactly our intuition behind using regularization.

The only difference between the JAMESSTEIN and ORACLEPLUGIN is that ORACLEPLUGIN contains an additional $d - 2$ term in the denominator, which gives ORACLEPLUGIN the extra property that it always returns an estimate between 0 and \bar{X} , whereas JAMESSTEIN can overshoot zero. Empirically, we found that ORACLEPLUGIN generally had a lower expected risk than JAMESSTEIN when $\|\theta_\infty\|$ is large (several values for θ_∞ , d and n were tested), but JAMESSTEIN was better when $\|\theta_\infty\| \leq 1$. For noisy settings, the standard James-Stein estimator is generally better.

One advantage of the analysis of the James-Stein estimator is that the exact expected risk for all n can be computed. Having this explicit form relies on being able to work with closed-form expressions for various expectations of Gaussian variables. By using asymptotics, we are able to side-step these specific computations and generalize to a much larger class of models, obtaining Stein’s result asymptotically as a special case.

Also note that we cannot obtain the James-Stein estimator exactly with a quadratic regularizer because the estimator resulting from the latter will have the form $\frac{\lambda}{n \|\bar{X}\|^2 + \lambda}$, not $\frac{\lambda}{n \|\bar{X}\|^2}$, as in James-Stein. James-Stein is only optimal over estimators of the form $\frac{\lambda}{n \|\bar{X}\|^2}$.

3.2 Binomial estimation

Setup We consider the estimation of the log-odds $\theta \in \mathbb{R}$ of a coin coming up heads given n i.i.d. coin flips. Then the negative log-likelihood loss corresponding to this probability model is

$$\ell(x; \theta) = -x\theta + \log(1 + e^\theta), \quad (46)$$

where $x \in \{0, 1\}$ is the outcome of the coin. One can verify that the probability of heads is $e^{-\ell(1; \theta)} = \frac{e^\theta}{1+e^\theta}$ and the probability of tails is $e^{-\ell(0; \theta)} = \frac{1}{1+e^\theta}$.

The main technical novelty in binomial estimation is that it requires reasoning about the asymptotic bias of the unregularized estimator (\mathcal{B}) appearing in (6), which is typically ignored in first-order asymptotics or is zero, as we saw in Gaussian mean estimation. Also note that in binomial estimation, the model is always well-specified, so we can assume $\mathcal{I}_{\ell\ell} = \ddot{\mathcal{L}}$ without loss of generality.

We consider the *conjugate regularizer* defined by:

$$R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta), \quad r(\theta) = -\frac{1}{2}\theta + \log(1 + e^\theta), \quad (47)$$

which corresponds to the negative log prior of a Beta($\frac{\lambda}{2} + 1, \frac{\lambda}{2} + 1$) distribution. Note that $r(\theta)$ has the same form as $\ell(\cdot, \theta)$ because the Beta distribution is conjugate to the binomial. Choosing λ (the hyperparameter of the Beta prior) has been studied extensively in statistics. Some common choices are the Haldane prior ($\lambda = -2$), the reference (Jeffreys) prior ($\lambda = -1$), the uniform prior ($\lambda = 0$), and Laplace smoothing ($\lambda = 2$). Our approach allows λ to be chosen to minimize expected risk adaptively based on data.

Oracle regularization To compute the oracle regularizer (10), we need to compute the various derivatives of the loss and regularizer. Because the binomial distribution is an exponential family, these derivatives correspond to the moments of the distribution, which we compute as follows:

$$\mu \stackrel{\text{def}}{=} \mathbb{E}_{\theta_\infty}[X] = \frac{1}{1 + e^{-\theta_\infty}}, \quad v \stackrel{\text{def}}{=} \text{Var}_{\theta_\infty}[X] = \mu(1 - \mu), \quad b \stackrel{\text{def}}{=} \mu - \frac{1}{2}, \quad (48)$$

where μ is the mean, v is the variance, and b is the offset from $\frac{1}{2}$. Remember that these quantities depend on θ_∞ , though we will suppress the dependence from the notation.

Now, the appropriate quantities in (10) can be computed: $\ddot{\mathcal{L}} = v$, $\ddot{\mathcal{L}} = -2vb$, $\dot{r} = b$, $\ddot{r} = v$. For the asymptotic bias, we have $\mathcal{I}_{\ell^2\ell} = v\dot{\mathcal{L}} = 0$ and $\mathcal{I}_{\ell\ell} = \ddot{\mathcal{L}}$, so $\mathcal{B} = v^{-1}b$. Note that \mathcal{B} tends to infinity as μ tends to 0 or 1. Indeed, maximum likelihood for binomial problems tends to underestimate the probability of rare events.

Plugging these quantities into (10), we get that the oracle regularization strength is

$$\lambda^* = \frac{1 + 2v^{-1}b^2}{v^{-1}b^2} = vb^{-2} + 2. \quad (49)$$

Note that λ^* is always positive, so (positive) regularization is always helpful. In particular, the minimum regularization strength is $\lambda^* = 2$ when $v = 0$ (corresponding to when the probability of heads μ is 0 or 1); $\lambda^* = 2$ corresponds to using Laplace smoothing. When $\mu \rightarrow \frac{1}{2}$, the oracle regularization strength tends to infinity.

We can calculate the corresponding relative risk:

$$\mathbb{L}(\lambda^*) = \frac{-(1 + 2v^{-1}b^2)^2}{2v^{-1}b^2} = -\left(\frac{1}{2}vb^{-2} + 2 + 2v^{-1}b^2\right). \quad (50)$$

Note that there are two cases when regularization helps immensely ($\mathbb{L}(\lambda^*) \rightarrow -\infty$): (1) when $\mu \rightarrow \frac{1}{2}$, we regularize heavily to stabilize the estimates; and (2) when μ tends to 0 or 1, we regularize with the minimum value of 2, but this also helps enormously in guarding against highly skewed estimates.

Plugin regularization Recall that PLUGIN depends on a function linking parameters to oracle regularization strengths:

$$f(\theta_\infty) = 2 + vb^{-2}. \quad (51)$$

Take its derivative:

$$\dot{f}(\theta_\infty) = (-2vb)b^{-2} + v(-2b^{-3})v = -2vb^{-1} - 2v^2b^{-3}. \quad (52)$$

Now we can compute difference in relative risk between PLUGIN and ORACLE:

$$\mathbb{L}^\bullet(1) - \mathbb{L}(\lambda^*) = \mathcal{E} = -b\dot{f}v^{-1} = 2 + 2vb^{-2}, \quad (53)$$

where we used the fact that $\nabla \dot{\mathcal{R}}(\lambda^*) = \dot{r} = b$. Since $\mathcal{E} \geq 2$, suboptimality of PLUGIN relative to ORACLE costs us asymptotically at least 2, with larger relative risk if μ is close to $\frac{1}{2}$.

Combining (50) and (53), we get:

$$\mathbb{L}^\bullet(1) = \frac{3}{2}vb^{-2} - 2v^{-1}b^2. \quad (54)$$

Therefore, PLUGIN is better than UNREGULARIZED when $\mathbb{L}^\bullet(1) < 0$, which happens when $3v^2 < 4b^4$; in terms of the heads probability, $\mu \geq 0.84$ or $\mu \leq 0.16$. Intuitively, we need the variance to be small enough so that the plugin estimates are reliable enough. When μ is close to $\frac{1}{2}$, regularization is not as critical in the first place, unless we can pinpoint λ^* , but estimating λ^* is difficult in this regime because the variance v is large.

3.3 Entropy regularization

Setup In prediction tasks, we wish to learn a mapping from some input $x \in \mathcal{X}$ to an output $y \in \mathcal{Y}$. A common approach is to use conditional exponential families, which are defined by a vector of sufficient statistics (features) $\phi(x, y) \in \mathbb{R}^d$ and an accompanying vector of parameters $\theta \in \mathbb{R}^d$ in the following way:

$$p_\theta(y | x) = \exp\{\phi(x, y)^\top \theta - A(\theta; x)\}, \quad A(\theta; x) = \log \int_{\mathcal{Y}} \exp\{\phi(x, y)^\top \theta\} dy. \quad (55)$$

To map this setup onto our general notation, let $z = (x, y)$ and $\ell(z; \theta) = -\log p_\theta(y | x)$. Conditional exponential families contain logistic regression and conditional random fields [22] as special cases.

Suppose in addition to our n labeled examples, we have m unlabeled examples X_{n+1}, \dots, X_{n+m} . We would like to exploit this unlabeled data to perform semi-supervised learning. One method of using the unlabeled data is to use entropy regularization [18], which *minimizes* the entropy of a learned distribution $p_\theta(y | x)$ as measured on unlabeled points. The motivation behind this criteria is that often in classification problems, the classes are well separated, and separation corresponds to having low entropy $p_\theta(y | x)$.

Formally, the *entropy regularizer* is defined as follows:

$$R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta), \quad r(\theta) = \frac{1}{m} \sum_{i=1}^m H(p_\theta(Y | X_{n+i})), \quad (56)$$

where $H(p(y | x)) = -\int p(y | x) \log p(y | x) dy$. Note that while R_n is a random quantity, it is independent of the labeled data.

Asymptotic analysis Now we study the asymptotic effects of entropy regularization. We assume that $m \rightarrow \infty$ as $n \rightarrow \infty$. First, we can compute the Hessian of the loss using standard moment-generating properties of the log-partition function $A(\theta; x)$:

$$\ddot{\mathcal{L}} = \mathbb{E}_{p^*(X)}[\mathbb{V}_{p_{\theta_\infty}(Y|X)}[\phi(X, Y) | X]], \quad (57)$$

Now we turn to the regularizer. Since $m \rightarrow \infty$, we have

$$\mathcal{R}(\lambda, \theta) = \lambda \cdot \mathbb{E}_{p^*(X)}[H(p_{\theta_\infty}(Y | X))]. \quad (58)$$

First, let us compute the derivative of the entropy function:

$$\nabla_\theta[H(p_\theta(Y | x))] = - \int \nabla_\theta[p_\theta(y | x) \log p_\theta(y | x)] dy \quad (59)$$

$$= - \int \nabla_\theta[p_\theta(y | x)] \log p_\theta(y | x) + p_\theta(y | x) \frac{\nabla_\theta[p_\theta(y | x)]}{p_\theta(y | x)} dy \quad (60)$$

$$= - \int \nabla_\theta[p_\theta(y | x)] (\log p_\theta(y | x) + 1) dy \quad (61)$$

$$= - \int p_\theta(y | x) [\phi(x, y) - \mathbb{E}_{p_\theta(Y|x)}[\phi(x, Y)]] (\phi(x, y)^\top \theta - A(\theta; x) + 1) dy \quad (62)$$

$$= - \int p_\theta(y | x) [\phi(x, y) - \mathbb{E}_{p_\theta(Y|x)}[\phi(x, Y)]] \phi(x, y)^\top \theta dy \quad (63)$$

$$= - \mathbb{V}_{p_\theta(Y|x)}[\phi(x, Y)] \theta \quad (64)$$

Define the expected conditional variance of the exponential family:

$$\mathcal{V}_x \stackrel{\text{def}}{=} \mathbb{E}_{p^*(X)}[\mathbb{V}_{p_{\theta_\infty}(Y|X)}[\phi(X, Y)]]. \quad (65)$$

From the derivative, it is clear that we reduce the entropy by increasing the magnitude of θ , especially along directions with high variance. Using this calculation, we obtain the derivatives for the loss and regularizer:

$$\ddot{\mathcal{L}} = \mathcal{V}_x, \quad (66)$$

$$\dot{\mathcal{R}} = -\mathcal{V}_x \theta_\infty, \quad (67)$$

$$\ddot{\mathcal{R}} = -\dot{\mathcal{V}}_x[\theta_\infty] - \mathcal{V}_x. \quad (68)$$

The oracle regularization strength can be computed:

$$\lambda^* = \frac{-\text{tr}\{\mathcal{I}_{\ell\ell} \mathcal{V}_x^{-1} (\dot{\mathcal{V}}_x[\theta_\infty] + \mathcal{V}_x) \mathcal{V}_x^{-1}\} - 2\mathcal{B}^\top \mathcal{V}_x \theta_\infty}{\theta_\infty^\top \mathcal{V}_x \theta_\infty}. \quad (69)$$

The sign and magnitude of λ^* provide some indication of the regimes in which entropy regularization should be helpful (large positive values indicate that regularization is useful). Unlike the regularizers we have considered so far, the entropy regularizer is non-convex, the variance reduction term is not necessarily positive. Therefore a small positive λ^* is not guaranteed to reduce the relative risk.

To gain some intuition, consider the case where \mathcal{X} is a singleton and $\mathcal{Y} = \{0, 1\}$, which brings us back to the task of binomial estimation. In this case, the oracle regularization strength simplifies to

$$\lambda^* = \frac{-(-2vb\theta_\infty + v)v^{-1} - 2(v^{-1}b)v\theta_\infty}{\theta_\infty^2 v} = -\frac{1}{\theta_\infty^2}. \quad (70)$$

Note that this quantity is negative, which means that (positive) entropy regularization always hurts performance. This is not surprising given that entropy regularization is in some sense ‘‘anti-regularization’’, pushing the parameter estimate μ towards 0 and 1 rather than shrinking towards $\frac{1}{2}$. For the case of a non-singleton input space \mathcal{X} (e.g., in logistic regression), we do not have such a simple and interpretable formula that captures the oracle regularization, but we suspect that the qualitative conclusion is similar to that of binomial estimation.

Does this analysis show that entropy regularization is not helpful? This is not necessarily the case, for our analysis comes with two caveats. First, this analysis is only an asymptotic one where the problem size is fixed and the sample size grows. Therefore, the regime in which the analysis applies is one in which we are trying to stabilize the estimator; however, in practice, we might be in the regime where the entropy regularizer is playing more of a global structural role in reducing the space of possible θ rather than refining the estimate of θ locally.

Second, we have been assuming that log-loss is the desired loss. Therefore, one pays a heavy penalty if one does not faithfully represent very low probability outcomes. As a result, regularization to place sufficient support on those outcomes is very important. On the other hand, in classification tasks in practice, 0-1 loss is often the preferred loss, in which case low probability events can be essentially ignored without sacrificing much performance. Entropy regularization seems to be tailored more to the coarser 0-1 loss.

3.4 Hybrid generative-discriminative learning

In this section, we are again concerned in prediction tasks, in which we would like to learn a mapping from an input $x \in \mathcal{X}$ to an output $y \in \mathcal{Y}$. Although one could solve this problem directly by learning a discriminative predictor of y given x , both theory and practice have demonstrated that one can perform better on prediction, especially for smaller training sets, by exploiting a generative model $p_\theta(x, y)$. In recent years, there has been interest in combining generative and discriminative learning [10, 30, 23, 27, 25].

Setup We consider generative and discriminative models defined by exponential families, where we let $\phi(x, y) \in \mathbb{R}^d$ denote the vector of sufficient statistics (features) of the exponential family and let $\theta \in \mathbb{R}^d$ be the parameters. These features and parameters can be used to define a generative model (71) or a discriminative model (72):

$$p_\theta(x, y) = \exp\{\phi(x, y)^\top \theta - A(\theta)\}, \quad A(\theta) = \log \int_{\mathcal{X}} \int_{\mathcal{Y}} \exp\{\phi(x, y)^\top \theta\} dy dx, \quad (71)$$

$$p_\theta(y | x) = \exp\{\phi(x, y)^\top \theta - A(\theta; x)\}, \quad A(\theta; x) = \log \int_{\mathcal{Y}} \exp\{\phi(x, y)^\top \theta\} dy. \quad (72)$$

Maximum (conditional) likelihood in these models leads to a generative estimator $\hat{\theta}_n^{\text{gen}}$ or a discriminative estimator $\hat{\theta}_n^{\text{dis}}$:

$$\hat{\theta}_n^{\text{gen}} \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} G_n(\theta), \quad G_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log p_\theta(x, y), \quad (73)$$

$$\hat{\theta}_n^{\text{dis}} \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} D_n(\theta), \quad D_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log p_\theta(y | x). \quad (74)$$

Since we are interested in prediction, our loss function is the negative log likelihood of the discriminative model, that is, $\ell(x, y) = -\log p_\theta(y | x)$.

The work of [25] showed that if the generative model is well-specified ($p^*(x, y) = p_{\theta_\infty}(x, y)$), then the generative estimator is better in the sense that

$$\mathcal{L}(\hat{\theta}_n^{\text{gen}}) \leq \mathcal{L}(\hat{\theta}_n^{\text{dis}}) - \frac{c}{n} + O_p(n^{-\frac{3}{2}}) \quad (75)$$

for some $c \geq 0$. This is because the generative model, by virtue of modeling x , has higher Fisher information than the discriminative model. On the other hand, if the model is misspecified, the discriminative estimator is asymptotically better, because the generative model does not even converge to the optimal limiting parameters θ_∞ .

We would like to refine the result of [25] by creating a hybrid estimator that interpolates between the generative and the discriminative estimators, so that if the model is “extremely mis-specified,” we can favor

the discriminative model, and if the model is “almost well-specified,” we give more weight to the generative model.

To formalize these intuitions, define a *hybrid estimator* in our regularization framework by treating the discriminative and generative objectives as the empirical risk and the regularizer, respectively:

$$\ell((x, y); \theta) = -\log p_\theta(y | x), \quad (76)$$

$$L_n(\theta) = D_n(\theta), \quad (77)$$

$$R_n(\lambda, \theta) = \frac{\lambda}{n} G_n(\theta). \quad (78)$$

Note that the regularizer is random as it depends on the training data. Our hybrid estimator is then just the standard regularized estimator:

$$\hat{\theta}_n^\lambda = \operatorname{argmin}_{\theta \in \mathbb{R}^d} D_n(\theta) + \frac{\lambda}{n} G_n(\theta). \quad (79)$$

Note that as $n \rightarrow \infty$, the discriminative objective $D_n(\theta)$ dominates as desired.

Asymptotic analysis Now we analyze the relative risk of these estimators and try to find the best value of λ . Our approach generalizes the analysis of [9], which applies only to unbiased estimators for conditionally well-specified models.

We need to compute the derivatives of the loss and regularizer. First, define the following moments of the generative and discriminative models:

$$\mu_{xy} \stackrel{\text{def}}{=} \mathbb{E}_{p^*(X, Y)}[\phi(X, Y)], \quad (80)$$

$$\mu_x \stackrel{\text{def}}{=} \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X, Y)], \quad (81)$$

$$\mu \stackrel{\text{def}}{=} \mathbb{E}_{p_{\theta_\infty}(X, Y)}[\phi(X, Y)], \quad (82)$$

$$\mathcal{V}_x \stackrel{\text{def}}{=} \mathbb{E}_{p^*(X)}[\mathbb{V}_{p_{\theta_\infty}(Y|X)}[\phi(X, Y)]], \quad (83)$$

$$\mathcal{V} \stackrel{\text{def}}{=} \mathbb{V}_{p_{\theta_\infty}(X, Y)}[\phi(X, Y)]. \quad (84)$$

By moment-generating properties of the exponential family, we have:

$$\ddot{\mathcal{L}} = \mathcal{V}_x, \quad (85)$$

$$\dot{\mathcal{R}}(\lambda) = \lambda \dot{r}, \quad \dot{r} = \mu - \mu_{xy}, \quad (86)$$

$$\ddot{\mathcal{R}}(\lambda) = \lambda \ddot{r}, \quad \ddot{r} = \mathcal{V}. \quad (87)$$

The oracle regularization strength is then

$$\lambda^* = \frac{\operatorname{tr}\{\mathcal{I}_{\ell\ell} \mathcal{V}_x^{-1} \mathcal{V} \mathcal{V}_x^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{xy}) - \operatorname{tr}\{\mathcal{I}_{\ell r} \mathcal{V}_x^{-1}\}}{\operatorname{tr}\{(\mu - \mu_{xy}) \otimes \mathcal{V}_x^{-1}\}}. \quad (88)$$

Well-specified discriminative model To gain more insight into (88), let us consider the simplified setting where the discriminative model is well-specified, that is, $p^*(y | x) = p_{\theta_\infty}(y | x)$. Note that this is a much weaker assumption than assuming the generative model is well-specified.

In this setting, we have $\mathcal{I}_{\ell\ell} = \ddot{\mathcal{L}}$ by the first Bartlett identity (Proposition 2). Next, we compute:

$$\mathcal{I}_{\ell r} = \mathbb{E}[(\phi(X, Y) - \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X, Y)])(\phi(X, Y) - \mathbb{E}_{p_{\theta_\infty}(X, Y)}[\phi(X, Y)])^\top] \quad (89)$$

$$= \mathbb{E}[(\phi(X, Y) - \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X, Y)])\phi(X, Y)^\top] \quad (90)$$

$$= \mathbb{E}[(\phi(X, Y) - \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X, Y)])(\phi(X, Y) - \mathbb{E}_{p^*(X)p_{\theta_\infty}(Y|X)}[\phi(X, Y)])^\top] \quad (91)$$

$$= \mathcal{V}_x. \quad (92)$$

Misspecification	$\text{tr}\{\mathcal{I}_{\ell\ell}\mathcal{V}_x^{-1}\mathcal{V}\mathcal{V}_x^{-1}\}$	$2\mathcal{B}^\top(\mu - \mu_{xy})$	$\text{tr}\{(\mu - \mu_{xy})^{\otimes 2}\mathcal{V}_x^{-1}\}$	λ^*	$\mathbb{L}(\lambda^*)$
0%	5	0	0	∞	-0.65
5%	5.38	-0.073	0.00098	310	-48
50%	13.8	-1.0	0.034	230	-808

Table 3: The oracle regularizer for the hybrid generative-discriminative estimator. As misspecification increases, we regularize less, but the relative risk is reduced more (due to more variance reduction).

Replacing $\mathcal{I}_{\ell\ell}$ and $\mathcal{I}_{\ell r}$ with our new expressions in (88), we get that the oracle regularization strength is

$$\lambda^* = \frac{\text{tr}\{(\mathcal{V} - \mathcal{V}_x)\mathcal{V}_x^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{xy})}{\text{tr}\{(\mu - \mu_{xy})^{\otimes 2}\mathcal{V}_x^{-1}\}}. \quad (93)$$

Now we interpret λ^* . Recall that larger positive values of λ^* means that we ought to leverage the generative model more. First, observe that $\mathcal{V} \succeq \mathcal{V}_x$ —that is, the generative model has larger Fisher information than the discriminative model; this was the key fact used in [25]. This identity means that the first term of the numerator is always non-negative with its magnitude equal to the fraction of missing information provided by the generative model. The second term $2\mathcal{B}^\top(\mu - \mu_{xy})$, is opaque at this level of generality; it is unclear whether this term is positive or negative.

Finally, the denominator, which is always positive, affects the magnitude of the regularization. Recall our intuition that how much we leverage the generative model depends on how well-specified it is. The denominator formalizes an asymptotic notion of misspecification, namely $\text{tr}\{(\mu - \mu_{xy})^{\otimes 2}\mathcal{V}_x^{-1}\}$, which is the Mahalanobis distance between the moments of the generative model (μ) and the moments under the true distribution (μ_{xy}). In the extreme case when the generative model is well-specified, we have $\mu = \mu_{xy}$.

An empirical example To provide some concrete intuition, we investigated the oracle regularizer for a synthetic binary classification problem of predicting $y \in \{0, 1\}$ from $x \in \{0, 1\}^k$. We use features $\phi(x, y) = (\mathbb{I}[y = 0]x^\top, \mathbb{I}[y = 1]x^\top)^\top \in \mathbb{R}^{2k}$ to define the generative (Naive Bayes) and discriminative (logistic regression) models.

We use $k = 5$ and $\theta_\infty = (\frac{1}{10}, \dots, \frac{1}{10}, \frac{3}{10}, \dots, \frac{3}{10})^\top$. The data generating distribution is a mixture between a well specified generative model and a distribution that enforces $x_1 = \dots = x_k$ (which clearly violates the Naive Bayes assumption):

$$p^*(x, y) = (1 - \varepsilon)p_{\theta_\infty}(x, y) + \varepsilon p_{\theta_\infty}(y)p_{\theta_\infty}(x_1 | y)\mathbb{I}[x_1 = \dots = x_k]. \quad (94)$$

Note that The amount of misspecification is controlled by $0 \leq \varepsilon \leq 1$, the fraction of examples whose features are perfectly correlated.

Table 3 shows how the oracle regularizer changes with ε . As ε increases, λ^* decreases (we regularize less) as expected. But perhaps surprisingly, the relative risk is reduced with more misspecification; this is due to the fact that the variance reduction term increases and has a quadratic effect on $\mathbb{L}(\lambda^*)$. Note that this effect happens because we are changing the learning problem; if we were holding the learning problem fixed (p^*) but only changing the generative model, we would not have this effect.

Figure 2 shows the relative risk $\mathbb{L}_n(\lambda)$ for various values of λ . The vertical line corresponds to λ^* , which was computed numerically by sampling. Note that the minimum of the curves ($\text{argmin}_\lambda \mathbb{L}_n(\lambda)$), the desired quantity, is quite close to λ^* and approaches λ^* as n increases, which empirically justifies our asymptotic approximations.

Unlabeled data One of the key advantages of having a generative model is that we can leverage unlabeled examples by treating the outputs as latent variables which are marginalized out. Specifically, suppose we have m i.i.d. unlabeled examples $X_{n+1}, \dots, X_{n+m} \sim p^*(x)$, with $m \rightarrow \infty$ as $n \rightarrow \infty$. Define the unlabeled

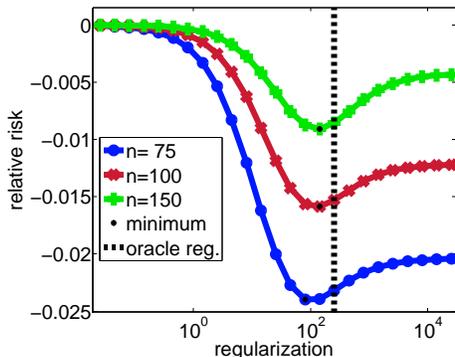


Figure 2: Relative risk $\mathbb{L}_n(\lambda)$ of the hybrid generative/discriminative estimator for various regularization strengths λ on a simple artificial example. Note that as the number of training points n increases, the λ attaining the minimum of $\mathbb{L}_n(\lambda)$ approaches the oracle λ^* (the dotted vertical line).

regularizer as

$$R_n(\lambda, \theta) = \frac{\lambda}{n} r(\theta), \quad r(\theta) = \frac{1}{m} \sum_{i=1}^m -\log p_\theta(X_{n+i}), \quad (95)$$

in a manner similar to entropy regularization (Section 3.3).

Using the fact that $\log p_\theta(x) = A(\theta; x) - A(\theta)$, we can differentiate the regularizer to obtain:

$$\dot{\mathcal{R}} = \mu - \mu_x, \quad (96)$$

$$\ddot{\mathcal{R}} = \mathcal{V} - \mathcal{V}_x. \quad (97)$$

Note that $\mathcal{I}_{\ell_r} = 0$, since the regularizer doesn't depend on the labeled data.

If the discriminative model is well-specified, then

$$\lambda^* = \frac{\text{tr}\{(\mathcal{V} - \mathcal{V}_x)\mathcal{V}_x^{-1}\} + 2\mathcal{B}^\top(\mu - \mu_{xy})}{\text{tr}\{(\mu - \mu_x) \otimes \mathcal{V}_x^{-1}\}}. \quad (98)$$

Since $\dot{\mathcal{L}}(\theta_\infty) = 0$, we have $\mu_x = \mu_{xy}$, so that our expression for λ^* is identical to the one from (93).

In summary, we see that regularizing the marginal likelihood on unlabeled data is asymptotically equivalent to regularizing the joint likelihood on labeled data if the discriminative model is well-specified. Although marginal likelihood has lower Fisher information than joint likelihood, the regularizer based on marginal likelihood uses unlabeled data independent of the labeled data; the joint likelihood has higher Fisher information, but is defined on the labeled data. These two factors balance out perfectly.

This equivalence suggests that the dominant asymptotic concern in hybrid generative-discriminative learning is developing a good generative model that is as well-specified as possible; the exact manner in which it is used in learning is less important. In practice, if one has a very small number of training points, regularizing on a large amount of unlabeled data is more advantageous than regularizing on the existing labeled data, but this difference is outside the scope of our asymptotic analysis.

3.5 Multi-task linear regression

Suppose that we want to solve K related tasks given n training examples per task. The idea of multi-task learning is to leverage the fact that the tasks are related and estimate the parameters for the tasks jointly, thereby sharing statistical strength between tasks. There have been a fair amount of work in the literature on multi-task learning (e.g., see [12, 3, 17, 2, 19] and reference therein).

Setup In this section, we focus on linear regression in the multi-task setting. We assume the following generative process for our data:

For each task $k = 1, \dots, K$:
 For each data point $i = 1, \dots, n$:
 Generate the input $X_i^k \sim p^*(X_i^k)$ where $\mathbb{E}[X_i^{k\otimes}] = I_d$
 Generate the output $Y_i^k \sim \mathcal{N}(X_i^{k\top} \theta_\infty^k, 1)$

By following this setup, we are assuming that the discriminative model is well-specified. The constraint $\mathbb{E}[X_i^{k\otimes}] = I_d$ on the true input distribution is mostly for convenience, as it leads to simpler expressions.

We can cast the multi-task problem as an ordinary single task problem by concatenating the vectors for all the tasks:

$$X_i = (X_i^{1\top}, \dots, X_i^{K\top})^\top \in \mathbb{R}^{Kd}, \quad Y = (Y^1, \dots, Y^K)^\top \in \mathbb{R}^K, \quad \theta = (\theta^{1\top}, \dots, \theta^{K\top})^\top \in \mathbb{R}^{Kd}. \quad (99)$$

It will also be useful to represent $\theta \in \mathbb{R}^{Kd}$ by the matrix $\Theta = (\theta^1, \dots, \theta^K) \in \mathbb{R}^{d \times K}$.

We use the standard squared loss function for each task. The loss function, defined over data points $(x, y) \in \mathbb{R}^{Kd} \times \mathbb{R}^K$, one from each task, is thus:

$$\ell((x, y), \theta) = \frac{1}{2} \sum_{k=1}^K (y^k - x^{k\top} \theta^k)^2. \quad (100)$$

The purpose of the regularizer is to shrink the parameters of the K tasks towards each other. In general, since some tasks are more related than others, it will be useful to allow the regularizer to shrink the tasks towards each other by varying amounts. To this end, let us define a positive definite matrix $\Lambda \in \mathbb{R}^{K \times K}$ of inter-task affinities. Define the quadratic regularizer as follows:

$$r(\Lambda, \theta) = \frac{1}{2} \theta^\top (\Lambda \otimes I_d) \theta = \text{tr}\{\Lambda \Theta^\top \Theta\}. \quad (101)$$

Note that there are $O(K^2)$ regularization parameters contained in the matrix Λ to be set.

Oracle regularization The derivation of the oracle regularizer closely parallels that of Gaussian mean estimation Section 3.1, but now extended to matrices. Let us first compute the various derivatives of the loss and regularizer:

$$\ddot{\mathcal{L}} = I_d \quad (102)$$

$$\ddot{\mathcal{R}} = (\Lambda \otimes I_d) \theta \quad (103)$$

$$\ddot{\mathcal{R}} = \Lambda \otimes I_d \quad (104)$$

Substituting these quantities into (6), we get the following expression for the relative risk:

$$\mathbb{L}(\Lambda) = \frac{1}{2} \text{tr}\{(\Lambda \otimes I_d)^2 \theta_\infty^{\otimes 2}\} - \text{tr}\{\Lambda \otimes I_d\} \quad (105)$$

$$= \frac{1}{2} \text{tr}\{\Lambda^2 \Theta_\infty^\top \Theta_\infty\} - d \cdot \text{tr}\{\Lambda\}, \quad (106)$$

where we used the fact that $\mathcal{I}_{\ell\ell} = \ddot{\mathcal{L}}$ because the discriminative model is well-specified.

Optimizing the relative risk $\mathbb{L}(\lambda)$ with respect to Λ produces the oracle regularization parameter:

$$\Lambda^* = d(\Theta_\infty^\top \Theta_\infty)^{-1}. \quad (107)$$

The associated relative risk is

$$\mathbb{L}(\Lambda^*) = -\frac{1}{2} d^2 \cdot \text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}. \quad (108)$$

Plugin regularization Now we analyze PLUGIN. Letting $f(\Theta) = d(\Theta^\top \Theta)^{-1}$, we compute its derivative:

$$\dot{f}(\Theta) = -d(\Theta^\top \Theta)^{-1}(2\Theta^\top (\cdot))(\Theta^\top \Theta)^{-1}. \quad (109)$$

Plugging this expression into (22), we get

$$\mathcal{E} = 2d \cdot \text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}. \quad (110)$$

Since $\mathbb{L}^\bullet(1) - \mathbb{L}(\lambda^*) = \mathcal{E}$, the asymptotic relative risk of PLUGIN is greater than that of ORACLE by this amount. Combining (108) and (110), we get that the asymptotic relative risk of PLUGIN is

$$\mathbb{L}^\bullet(1) = -\frac{1}{2}d(d-4)\text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}. \quad (111)$$

When this quantity is negative, PLUGIN is better than UNREGULARIZED; this happens for $d > 4$.

Oracle plugin regularization Now let us derive ORACLEPLUGIN. By Theorem 3, we get that the plugin regularizer just needs to be reweighted by:

$$\lambda^{\bullet*} = 1 + \frac{\mathcal{E}}{2\mathbb{L}(\lambda^*)} = \frac{2d \cdot \text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}}{-d^2 \cdot \text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}} = 1 - \frac{2}{d}. \quad (112)$$

The asymptotic relative risk of ORACLEPLUGIN is

$$\mathbb{L}^\bullet(\lambda^{\bullet*}) = \mathbb{L}^\bullet(1) + \frac{\mathcal{E}^2}{4\mathbb{L}(\lambda^*)} = -\frac{1}{2}(d-2)^2\text{tr}\{(\Theta_\infty^\top \Theta_\infty)^{-1}\}. \quad (113)$$

When this quantity is negative, ORACLEPLUGIN is better than UNREGULARIZED; this happens for $d > 2$.

Joint versus independent regularization A principal question is how much we improve performance by using multi-task learning rather than learning the parameters of each task independently. If we solve the K regression tasks independently with K independent regularization parameters (which are set according to ORACLEPLUGIN), the asymptotic relative risk would be

$$\mathbb{L}_{\text{indep}}^\bullet(\lambda^{\bullet*}) = -\frac{1}{2}(d-2)^2 \sum_{k=1}^K \|\theta_\infty^k\|^{-2}, \quad (114)$$

since the risks are simply additive over the tasks.

Now let us compare the relative risks of ORACLEPLUGIN using joint versus independent regularization. Let $A = \Theta_\infty^\top \Theta_\infty$ with eigendecomposition $A = UDU^\top$. Then the asymptotic relative risk of joint regularization can be written as

$$\mathbb{L}^\bullet(\lambda^{\bullet*}) = -\frac{1}{2}(d-2)^2 \sum_{k=1}^K D_{kk}^{-1}, \quad (115)$$

since $\text{tr}\{A^{-1}\} = \text{tr}\{D^{-1}\}$. The asymptotic relative risk of independent regularization can be written as

$$\mathbb{L}_{\text{indep}}^\bullet(\lambda^{\bullet*}) = -\frac{1}{2}(d-2)^2 \sum_{k=1}^K A_{kk}^{-1}. \quad (116)$$

The gap between joint and independent regularization is large when the tasks are non-trivial but similar (θ_∞^k s are close, but $\|\theta_\infty^k\|$ is fairly large). In this case, the first eigenvalue of A is large (corresponding to the principal direction along which the θ_∞^k s point) and the rest are small. Therefore, D_{kk}^{-1} is small for $k = 1$ but quite large for $k > 1$. Therefore, $\mathbb{L}^\bullet(\lambda^{\bullet*})$ is favorable. In contrast, $A_{kk}^{-1} = \|\theta_\infty^k\|^2$ is small for all k , and therefore, $\mathbb{L}_{\text{indep}}^\bullet(\lambda^{\bullet*})$ is relatively small in magnitude.

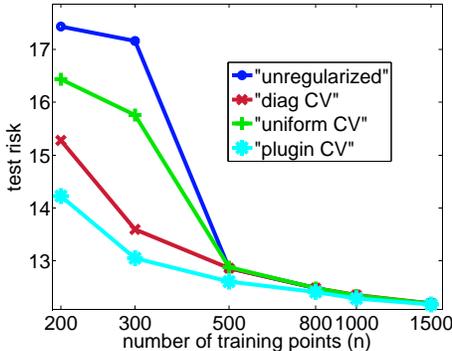


Figure 3: On the MHC-I binding prediction task, test risk for the four multi-task estimators. See text for the description of the estimators. PLUGINCV (estimating all pairwise task affinities using PLUGIN and cross-validating the strength) works best.

MHC-I binding prediction We evaluated our multi-task regularization method on the IEDB MHC-I peptide binding dataset created by [29] and used by [19]. The goal here is to predict the binding affinity (represented by $\log IC_{50}$) of a MHC-I molecule given its amino-acid sequence (represented by a vector of binary features, reduced to a 20-dimensional real vector using SVD). We created five regression tasks corresponding to the five most common MHC-I molecules.

We compared four estimators:

- UNREGULARIZED: no regularization was used;
- DIAGCV ($\Lambda = cI$): solve each regression task independently;
- UNIFORMCV ($\Lambda = c(\mathbf{1}\mathbf{1}^\top + 10^{-5}I_K)$): use a multi-task regularizer that shrinks the tasks towards each other, but using the same task-affinity for all pairs of tasks; and
- PLUGINCV ($\Lambda = cd(\hat{\Theta}_n^\top \hat{\Theta}_n)^{-1}$): using the PLUGIN estimator with regularization strength c .

In all cases, the parameter where c was chosen by three-fold cross-validation from 21 candidates in $[10^{-5}, 10^5]$.

Figure 3 shows the results averaged over 30 independent train/test splits. First, note that if we assume all tasks are equally related (UNIFORMCV), multi-task regularization actually performs worse than independent learning (DIAGCV). By learning the full matrix of task affinities (PLUGINCV), we obtain the best results.

We also experimented with setting c directly via ORACLEPLUGIN rather than cross-validation; this did not work very well, presumably to the inexactness of the asymptotics. Nonetheless, our asymptotic analysis is useful for producing the form of the regularizer (determined by Λ) by setting the $O(K^2)$ regularization parameters. This would have not been computational feasible to do via cross-validation. Note that there are alternative approaches for optimizing multiple hyperparameters for multi-task learning [19].

4 Related work and discussion

The problem of choosing regularizers has received much attention in both the machine learning literature and the statistics literature [8]. In this section, we attempt to place our asymptotic analysis of regularization in relation to existing work.

Relationship to learning theory bounds In machine learning, much of the learning theory literature focuses on risk bounds, which approximate the expected risk ($\mathcal{L}(\hat{\theta}_n^\lambda)$) with upper bounds. Our asymptotic analysis provides a different type of approximation—one that is exact in the first few terms of the expansion of the risk, but also one that makes no precise statement about the risk for any fixed n .

The two approaches also deal with fundamentally different aspects of the problem: Risk bounds are generally based on the complexity of the hypothesis class, whereas asymptotic expansions are based on the variance of the estimator. [26] shows that these analyses are complementary and actually capture different regimes of the learning curve.

Vanilla uniform convergence bounds yield worst-case analyses, whereas our asymptotic analysis is tailored to a particular problem (p^* and θ_∞) and algorithm (estimator). Localization techniques [5], regret analyses [13], and stability-based bounds [11] all allow for some degree of problem- and algorithm-dependence. As bounds, however, they necessarily have some looseness, whereas our analysis provides exact constants, at least the ones associated with the lowest-order terms.

One of the principal advantages of using asymptotic expansions rather than bounds is that exact control over the first few terms allows us to, at least asymptotically, compare the quality of different estimators and choose the best one.

Relationship to asymptotic analyses in statistics Asymptotics has a rich tradition in statistics, in particular for model selection. One of the classic analyses is the Akaike information criterion (AIC) [1], which approximates the risk of the maximum likelihood estimate by subtracting off the asymptotic bias. There have been many extensions to AIC [32, 31, 28, 21, 16] which apply in the misspecified setting and work for estimators other than maximum likelihood.

Note that though our goal is different—choosing regularization parameters rather than performing model selection. However, we share the general idea of performing an asymptotic expansion of the risk. An additional difference is that we are working with a single parametric model, for which we know that maximum likelihood is asymptotically efficient—that is, obtains the minimum asymptotic risk. Therefore, we need to consider second-order asymptotics to elicit the differences in the relative risk across different regularization parameters. Higher-order asymptotics has been used in many statistical settings, for example, in the context of estimating means [6, 24] or estimating shift parameters in semiparametric models [15].

Last remarks Another method for choosing regularization parameters, which is perhaps the most common in practice, is cross-validation [14]. However, note that cross-validation is feasible only when the number of regularization parameters is very small, as the optimization of the regularization parameters often can only be approached in a black-box manner. In contrast, our approach can optimize many hyperparameters at once. Perhaps the most effective solution is to combine the two approaches by optimizing the form of the regularizer using our asymptotic analyses and further calibrating the regularization weight using cross-validation. This is a useful technique that we exploited in multi-task learning (Section 3.5).

To conclude, we have developed a general asymptotic framework for analyzing regularization, along with an efficient procedure for choosing regularization parameters based on asymptotic criteria. We believe that the tools we have developed provide a complementary perspective on analyzing learning algorithms to that of risk bounds, thus deepening our understanding of regularization. An important direction for future work is to develop analyses that (1) work for non-smooth losses and regularizers and (2) work in a non-parametric setting.

References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- [2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 41–48, 2007.

- [3] B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.
- [4] M. S. Bartlett. Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40:306–317, 1953.
- [5] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
- [6] P. E. Berkhin and B. Y. Levit. Second-order asymptotically minimax estimates for the mean of a normal population. *Problemy Peredachi Informatsii*, 16:60–79, 1980.
- [7] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistics Society: Series B (Statistical Methodology)*, 41:113–147, 1979.
- [8] P. Bickel and B. Li. Regularization in statistics. *Sociedad de Estadística e Investigación Operativa Test*, 15:271–344, 2006.
- [9] G. Bouchard. Bias-variance tradeoff in hybrid generative-discriminative models. In *Sixth International Conference on Machine Learning and Applications (ICMLA)*, pages 124–129, 2007.
- [10] G. Bouchard and B. Triggs. The trade-off between generative and discriminative classifiers. In *International Conference on Computational Statistics*, pages 721–728, 2004.
- [11] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- [12] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.
- [13] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- [14] P. Craven and G. Wahba. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1978.
- [15] A. S. Dalalyan, G. K. Golubev, and A. B. Tsybakov. Penalized maximum likelihood and semiparametric second-order efficiency. *Annals of Statistics*, 34(1):169–201, 2006.
- [16] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.
- [17] T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.
- [18] Y. Grandvalet and Y. Bengio. Entropy regularization. In *Semi-Supervised Learning*, 2005.
- [19] L. Jacob, F. Bach, and J. Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 745–752, 2009.
- [20] W. James and C. Stein. Estimation with quadratic loss. In *Fourth Berkeley Symposium in Mathematics, Statistics, and Probability*, pages 361–380, 1961.
- [21] S. Konishi and G. Kitagawa. Generalized information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.
- [22] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling data. In *International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

- [23] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 87–94, 2006.
- [24] B. Y. Levit. Second-order asymptotic optimality and positive solutions of the schrödinger equation. *Theory of Probability and its Applications*, 30:333–363, 1985.
- [25] P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning (ICML)*, 2008.
- [26] P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *International Conference on Machine Learning (ICML)*, 2010.
- [27] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2006.
- [28] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion—determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.
- [29] B. Peters, H. Bui, S. Frankild, M. Nielson, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harn-dahl, W. Fleri, S. S. Wilson, J. Sidney, O. Lund, S. Buus, and A. Sette. A community resource bench-marking predictions of peptide binding to MHC-I molecules. *PLoS Computational Biology*, 2, 2006.
- [30] R. Raina, Y. Shen, A. Ng, and A. McCallum. Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.
- [31] R. Shibata. Statistical aspects of model selection. In *From Data to Model*, pages 215–240. 1989.
- [32] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.
- [33] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A Proof of Theorem 1 and Theorem 2

In the following, we hold λ fixed, so we will omit it from the equations that follow. First define the training objective:

$$M_n(\theta) \stackrel{\text{def}}{=} L_n(\theta) + R_n(\theta). \tag{117}$$

Let

$$\delta_n \stackrel{\text{def}}{=} \hat{\theta}_n - \theta_\infty \tag{118}$$

denote the parameter error of the regularized estimator with the fixed λ .

It is clear that the regularizer should go to zero in order to obtain consistency ($\delta_n \xrightarrow{P} 0$). In other words, if $R_n(\theta) = O_p(a_n)$, then we require that $a_n \rightarrow 0$. But what rate should this convergence happen? Let us assume that a_n decays fast enough so that $\delta_n = O_p(n^{-\frac{1}{2}})$. This is always possible because $O_p(n^{-\frac{1}{2}})$ is achievable without regularization, and we really shouldn't make things worse with regularization. Also, we assume that \dot{R}_n and \ddot{R}_n are also $O_p(a_n)$.

The plan is as follows: We first derive an asymptotic expression for the parameter error δ_n . Then we plug this result into a Taylor expansion of the risk \mathcal{L} to get the final expression. From this, we will see that the optimal rate is $a_n = O(n^{-1})$.

A.1 Asymptotic parameter error

We assume that unregularized estimator is consistent, that is, $\hat{\theta}_n^0 \xrightarrow{P} \theta_\infty$. This licenses us to perform a Taylor expansion. Taylor-expand the derivative of the training objective around θ_∞ to the second-order (by assumption, $\delta_n = O_p(n^{-\frac{1}{2}})$):

$$0 = \dot{M}_n(\hat{\theta}_n) = \dot{M}_n + \ddot{M}_n \delta_n + \frac{1}{2} \ddot{\ddot{M}}_n [\delta_n^\otimes] + O_p(n^{-\frac{3}{2}}). \quad (119)$$

Perform some algebra on (119) to get:

$$\delta_n = -(\ddot{M}_n + \frac{1}{2} \ddot{\ddot{M}}_n [\delta_n] + O_p(n^{-1}))^{-1} \dot{M}_n. \quad (120)$$

Taylor-expand the inverse function around \ddot{M}_n :

$$\delta_n = -\ddot{M}_n^{-1} \dot{M}_n + \ddot{M}_n^{-1} \left(\frac{1}{2} \ddot{\ddot{M}}_n \delta_n + O_p(n^{-1}) \right) \ddot{M}_n^{-1} \dot{M}_n. \quad (121)$$

Note that this is a recursive definition of δ_n . To remove the recursion, note that the first-order expansion (first two terms of (119)) yields $\delta_n = -\ddot{M}_n^{-1} \dot{M}_n + O_p(n^{-1})$. Plugging this in for δ_n yields:

$$\delta_n = -\ddot{M}_n^{-1} \dot{M}_n - \ddot{M}_n^{-1} \left(\frac{1}{2} \ddot{\ddot{M}}_n [\ddot{M}_n^{-1} \dot{M}_n] + O_p(n^{-1}) \right) \ddot{M}_n^{-1} \dot{M}_n. \quad (122)$$

Simplifying and rearranging terms:

$$\delta_n = -\ddot{M}_n^{-1} \dot{M}_n - \frac{1}{2} \ddot{M}_n^{-1} \ddot{\ddot{M}}_n [(\ddot{M}_n^{-1} \dot{M}_n)^\otimes] + O_p(n^{-\frac{3}{2}}). \quad (123)$$

We have thus obtained the asymptotic expression for δ_n in terms of derivatives of the training objective $M_n(\theta)$. The next step will be to rewrite this expression explicitly in terms of (the derivatives of) L_n and R_n . The expansions of the relevant quantities are as follows:

$$\dot{M}_n = \dot{L}_n + \dot{R}_n, \quad (124)$$

$$\ddot{M}_n^{-1} = (\ddot{L}_n + \ddot{R}_n)^{-1} = \ddot{L}_n^{-1} - \ddot{L}_n^{-1} \ddot{R}_n \ddot{L}_n^{-1} + O_p(a_n^2), \quad (125)$$

$$\ddot{\ddot{M}}_n = \ddot{\ddot{L}}_n + \ddot{\ddot{R}}_n, \quad (126)$$

where the first and third equations follow from linearity and the second follows by Taylor expanding the inverse function around \ddot{L}_n , leveraging the fact that $\ddot{R}_n = O_p(a_n)$.

To simplify notation, define

$$\dot{U}_n \stackrel{\text{def}}{=} \ddot{L}_n^{-1} \dot{L}_n, \quad \ddot{U}_n \stackrel{\text{def}}{=} \ddot{L}_n^{-1} \ddot{\ddot{L}}_n, \quad (127)$$

$$\dot{V}_n \stackrel{\text{def}}{=} \ddot{L}_n^{-1} \dot{R}_n, \quad \ddot{V}_n \stackrel{\text{def}}{=} \ddot{L}_n^{-1} \ddot{R}_n. \quad (128)$$

We split the parameter error δ_n (123) into two parts, one that does not depend on the regularizer (B_n) and one that does (C_n):

$$\delta_n = B_n + C_n, \quad (129)$$

where

$$B_n \stackrel{\text{def}}{=} -\dot{U}_n - \frac{1}{2} \ddot{U}_n [\dot{U}_n^\otimes] + O_p(n^{-\frac{3}{2}}), \quad (130)$$

$$C_n \stackrel{\text{def}}{=} -\dot{V}_n + \ddot{V}_n \dot{U}_n - \ddot{U}_n [\dot{U}_n \dot{V}_n^\top] + O_p(a_n n^{-1}) + O_p(a_n^2). \quad (131)$$

These equations are obtained by expanding \dot{M}_n , \ddot{M}_n^{-1} , and $\ddot{\ddot{M}}_n$ in the context of (123) and separating out the cross terms, and only keeping the ones which are large enough.

A.2 Asymptotic risk

Having a handle on the parameters, we turn to the risk. Expand the risk:

$$\mathcal{L}(\hat{\theta}_n) = \mathcal{L} + \underbrace{\dot{\mathcal{L}}[\delta_n]}_{=0} + \frac{1}{2}\ddot{\mathcal{L}}[\delta_n^{\otimes 2}] + \frac{1}{6}\dddot{\mathcal{L}}[\delta_n^{\otimes 3}] + O_p(n^{-2}). \quad (132)$$

Note that $\dot{\mathcal{L}}(\theta_\infty) = 0$ because θ_∞ is the minimizer of \mathcal{L} . Now we will compute $\mathbb{L}_n = \mathbb{E}[\mathcal{L}(\hat{\theta}_n) - \mathcal{L}(\hat{\theta}_n^0)]$ from (132), which requires expanding $\delta_n^{\otimes 2}$ and $\delta_n^{\otimes 3}$ using (129). Before doing this computation, note the following:

- All terms that don't depend on the regularizer (that is, those terms that don't involve C_n) cancel in the relative risk \mathbb{L}_n , so we do not consider them.
- We absorb all terms of order $o(a_n n^{-1})$ or $o(a_n^2)$ into \dots .
- Recall that $\dot{U}_n = O_p(n^{-\frac{1}{2}})$, $\ddot{U}_n = O_p(1)$, $\dot{V}_n = O_p(a_n)$, and $\ddot{V}_n = O_p(a_n)$.
- We have $a_n^{-1}\dot{R}_n \xrightarrow{P} \dot{\mathcal{R}}$ for some $\dot{\mathcal{R}}$; $a_n\dot{R}_n \xrightarrow{P} \dot{\mathcal{R}}$ for some $\dot{\mathcal{R}}$; and $a_n^{-1}n\mathbb{E}[\dot{L}_n\dot{R}_n^\top] \rightarrow \mathcal{I}_{\ell r}$.
- We have $\ddot{L}_n^{-1} \xrightarrow{P} \ddot{\mathcal{L}}^{-1}$, and $n\dot{L}_n^{\otimes 2} \xrightarrow{P} \mathcal{I}_{\ell\ell}$.

Using these points, we compute and keep the following terms of the relative risk \mathbb{L}_n :

- (a) $\frac{1}{2}\ddot{\mathcal{L}}$ times the expectation of twice² the outer product of the first term in B_n and the first term in C_n :

$$\mathbb{E}[\ddot{\mathcal{L}}[\dot{U}_n\dot{V}_n^\top]] = \mathbb{E}[\ddot{\mathcal{L}}[\ddot{L}_n^{-1}\dot{L}_n\dot{R}_n^\top\ddot{L}_n^{-1}]]. \quad (133)$$

We expand this last expression by taking a multivariate expansion of \ddot{L}_n^{-1} around $\ddot{\mathcal{L}}^{-1}$, of \dot{L}_n around 0, and of \dot{R}_n around $\mathbb{E}[\dot{R}_n]$. Note that $\mathbb{E}[\dot{L}_n] = 0$, so all cross-terms that do not include the product of \dot{L}_n and another random variable vanish. Here are the terms which are left in the expansion:

$$\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}(\ddot{L}_n^{-1} - \ddot{\mathcal{L}}^{-1})\dot{L}_n\mathbb{E}[\dot{R}_n^\top]\ddot{\mathcal{L}}^{-1}\}] + \quad (134)$$

$$\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\ddot{L}_n^{-1}\dot{L}_n(\dot{R}_n - \mathbb{E}[\dot{R}_n])^\top\ddot{\mathcal{L}}^{-1}\}] + \quad (135)$$

$$\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\ddot{L}_n^{-1}\dot{L}_n\mathbb{E}\dot{R}_n^\top(\ddot{L}_n^{-1} - \ddot{\mathcal{L}}^{-1})\}] + o(a_n n^{-1}) \quad (136)$$

$$= 2\mathbb{E}[\ddot{L}_n^{-1}\dot{L}_n]^\top\mathbb{E}\dot{R}_n + \text{tr}\{\mathbb{E}[\dot{L}_n\dot{R}_n^\top]\ddot{\mathcal{L}}^{-1}\} + o(a_n n^{-1}) \quad (137)$$

$$= 2a_n\mathbb{E}[\dot{U}_n]^\top\dot{\mathcal{R}} + a_n n^{-1}\text{tr}\{\mathcal{I}_{\ell r}\ddot{\mathcal{L}}^{-1}\} + o(a_n n^{-1}). \quad (138)$$

- (b) $\frac{1}{2}\ddot{\mathcal{L}}$ times the expectation of twice the outer product of the first term of B_n and the second term of C_n :

$$-\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\ddot{V}_n\dot{U}_n^{\otimes 2}\}] = -a_n n^{-1}\text{tr}\{\mathcal{I}_{\ell\ell}\ddot{\mathcal{L}}^{-1}\ddot{\mathcal{R}}\ddot{\mathcal{L}}^{-1}\} + o(a_n n^{-1}).$$

- (c) $\frac{1}{2}\ddot{\mathcal{L}}$ times the expectation of twice the outer product of the first term of B_n and the third term of C_n :

$$\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\ddot{U}_n[\dot{U}_n\dot{V}_n^\top]\dot{U}_n^\top\}] = \mathbb{E}[\ddot{L}_n[\dot{U}_n^{\otimes 2} \otimes \dot{V}_n]] + o(a_n n^{-1}) = a_n\mathbb{E}[\ddot{U}_n[\dot{U}_n^{\otimes 2}]]^\top\dot{\mathcal{R}} + o(a_n n^{-1}).$$

- (d) $\frac{1}{2}\ddot{\mathcal{L}}$ times the expectation of twice the outer product of the second term of B_n and the first term of C_n :

$$\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\frac{1}{2}\ddot{U}_n[\dot{U}_n^{\otimes 2}]\dot{V}_n^\top\}] = \frac{1}{2}\mathbb{E}[\ddot{L}_n[\dot{U}_n^{\otimes 2} \otimes \dot{V}_n]] + o(a_n n^{-1}) = \frac{1}{2}a_n\mathbb{E}[\ddot{U}_n[\dot{U}_n^{\otimes 2}]]^\top\dot{\mathcal{R}} + o(a_n n^{-1}).$$

²Once for $B_n C_n^\top$ and once for $C_n B_n^\top$.

(e) $\frac{1}{2}\ddot{\mathcal{L}}$ times the expectation of twice the square of the first term of C_n :

$$\frac{1}{2}\mathbb{E}[\text{tr}\{\ddot{\mathcal{L}}\dot{V}_n^\otimes\}] = \frac{1}{2}a_n^2\text{tr}\{\dot{\mathcal{R}}^\otimes\ddot{\mathcal{L}}^{-1}\} + o(a_n^2).$$

(f) The $\frac{1}{6}\dddot{\mathcal{L}}$ times thrice the square of the first term of B_n times the first term of C_n :

$$-\frac{1}{2}\dddot{\mathcal{L}}[\dot{U}_n^\otimes \otimes \dot{V}_n] = -\frac{1}{2}a_n\mathbb{E}[\ddot{U}_n[\dot{U}_n^\otimes]]^\top \dot{\mathcal{R}} + o(a_n n^{-1}).$$

Note that the first term in (a) contains $\mathbb{E}[\dot{U}_n]$ and the sum of (c),(d), and (f) contains $\mathbb{E}[\ddot{U}_n[\dot{U}_n^\otimes]]$, which come together to create twice the expected parameter error $2\mathbb{E}[B_n]$. Note that B_n is also referred to as the unregularizer estimator bias, which has a limit according to $n\mathbb{E}B_n \rightarrow \mathcal{B}$. Using this information, we have:

$$\mathbb{L}_n = \underbrace{\frac{1}{2}a_n^2\text{tr}\{\dot{\mathcal{R}}^\otimes\ddot{\mathcal{L}}^{-1}\}}_{(e)} - \underbrace{a_n n^{-1}\text{tr}\{\mathcal{I}_{\ell\ell}\ddot{\mathcal{L}}^{-1}\dot{\mathcal{R}}\ddot{\mathcal{L}}^{-1}\}}_{(b)} - \underbrace{2a_n n^{-1}\mathcal{B}^\top \dot{\mathcal{R}}}_{(a),(c),(d),(f)} + \underbrace{a_n n^{-1}\text{tr}\{\mathcal{I}_{\ell r}\ddot{\mathcal{L}}^{-1}\}}_{(a)} + \dots \quad (139)$$

To minimize this quantity, we want the $O(a_n n^{-1})$ terms to balance the $O(a_n^2)$ terms, implying that $a_n = n^{-1}$ is in general the optimal rate. Substituting n^{-1} for a_n yields (6), thus completing the proof.

B Proof of Proposition 1

Expanding the asymptotic expansion of the parameter error (130) and taking expectations, we get that

$$\mathbb{E}[\hat{\theta}_n^0 - \theta_\infty] = -\mathbb{E}[\ddot{L}_n^{-1}\dot{L}_n] - \frac{1}{2}\mathbb{E}[\ddot{L}_n^{-1}\ddot{L}_n[\ddot{L}_n^{-1}\dot{L}_n^\otimes\ddot{L}_n^{-1}]] + O_p(n^{-\frac{3}{2}}). \quad (140)$$

Let's tackle the first term of the right-hand side of (140). Taylor expand \ddot{L}_n^{-1} :

$$\ddot{L}_n^{-1} = \ddot{\mathcal{L}}^{-1} - \ddot{\mathcal{L}}^{-1}(\ddot{L}_n - \ddot{\mathcal{L}})\ddot{\mathcal{L}}^{-1} + O_p(n^{-1}). \quad (141)$$

Plugging this into the first term of (140) yields:

$$-\mathbb{E}[\ddot{L}_n^{-1}\dot{L}_n] = -\mathbb{E}[\ddot{\mathcal{L}}^{-1}\dot{L}_n] + \mathbb{E}[\ddot{\mathcal{L}}^{-1}(\ddot{L}_n - \ddot{\mathcal{L}})\ddot{\mathcal{L}}^{-1}\dot{L}_n] + O_p(n^{-\frac{3}{2}}) \quad (142)$$

$$= \mathbb{E}[\ddot{\mathcal{L}}^{-1}\ddot{L}_n\ddot{\mathcal{L}}^{-1}\dot{L}_n] + O_p(n^{-\frac{3}{2}}), \quad (143)$$

where the second inequality uses the fact that $\mathbb{E}[\dot{L}_n] = 0$. Rearrange to put the random quantities next to each other, creating a random rank-3 tensor:

$$\mathbb{E}[\ddot{\mathcal{L}}^{-1}\ddot{L}_n\ddot{\mathcal{L}}^{-1}\dot{L}_n] = \mathbb{E}[\ddot{\mathcal{L}}^{-1}(\ddot{L}_n \otimes \dot{L}_n)[\ddot{\mathcal{L}}^{-1}]]. \quad (144)$$

Letting $n \rightarrow \infty$ and scaling by n yields $\ddot{\mathcal{L}}^{-1}\mathcal{I}_{\ell^2\ell}\ddot{\mathcal{L}}^{-1}$.

The second term of the right-hand side of (140) is simpler because all its factors have nonzero limiting expectations. Scaling by n (for \dot{L}_n^\otimes to converge to a non-trivial limit), we get $\ddot{\mathcal{L}}^{-1}\ddot{\mathcal{L}}[\ddot{\mathcal{L}}^{-1}\mathcal{I}_{\ell\ell}\ddot{\mathcal{L}}^{-1}]$.

C Proof of Proposition 2

In the well-specified setting, we have $\ell(z; \theta_\infty) = -\log p^*(z_2 | z_1)$. Start with the fact that probabilities integrate to 1:

$$\int e^{-\ell(z; \theta_\infty)} p^*(z_1) dz = \mathbb{E}[1] = 1. \quad (145)$$

With sufficient regularity conditions, differentiate the integrand of (145) with respect to θ :

$$\int e^{-\ell(z; \theta_\infty)} [-\dot{\ell}(z; \theta_\infty)] p^*(z_1) dz = 0. \quad (146)$$

Using the fact that $e^{-\ell(z; \theta_\infty)} p^*(z_1) = p^*(z_1, z_2)$, the left-hand side is equivalent to $\mathbb{E}[-\dot{\ell}(Z; \theta_\infty)]$. Therefore, we get the first Bartlett identity: $\dot{\mathcal{L}} = 0$. Note that $\dot{\mathcal{L}} = \dot{\mathcal{L}}(\theta_\infty) = 0$ actually holds regardless of whether the model is well-specified because θ_∞ is always chosen to minimize the expected loss $\mathcal{L}(\theta)$, which is attained when $\dot{\mathcal{L}}(\theta) = 0$.

Now differentiate (146) with respect to θ :

$$\int e^{-\ell(z; \theta_\infty)} [\dot{\ell}(z; \theta_\infty)^\otimes - \ddot{\ell}(z; \theta_\infty)] p^*(z_1) dz = 0. \quad (147)$$

Again, note the left-hand side is $\mathbb{E}[\dot{\ell}(Z; \theta_\infty)^\otimes - \ddot{\ell}(Z; \theta_\infty)]$, which is exactly $\mathcal{I}_{\ell\ell} - \dot{\mathcal{L}} = 0$.

Note that we can continue differentiating to obtain higher-order Bartlett identities, but these will not be useful to us.

D Proof of Theorem 3

Proof. Since $\hat{\lambda}_n \xrightarrow{P} \lambda^*$, $\dot{\mathcal{R}}^\bullet(\lambda^\bullet) = \lambda^\bullet \dot{\mathcal{R}}(\lambda^*)$ and $\ddot{\mathcal{R}}^\bullet(\lambda^\bullet) = \lambda^\bullet \ddot{\mathcal{R}}(\lambda^*)$. We can evaluate $\mathbb{L}_n^\bullet(\lambda^\bullet)$ using (6) as a reference. In particular, the first term is multiplied by $(\lambda^\bullet)^2$, the second and third terms are multiplied by λ^\bullet , and the last term changes in the following way:

$$\text{tr}\{\mathcal{I}_{\ell r}^\bullet \ddot{\mathcal{L}}^{-1}\} = \lim_{n \rightarrow \infty} n^2 \mathbb{E}[\text{tr}\{\dot{L}_n \lambda^\bullet \dot{R}_n(\lambda^* + \varepsilon_n)^\top \ddot{\mathcal{L}}^{-1}\}] = \lambda^\bullet (\text{tr}\{\mathcal{I}_{\ell r} \ddot{\mathcal{L}}^{-1}\} + \mathcal{E}),$$

where the first equality used the fact that $\dot{R}_n^\bullet(\lambda^\bullet) = \lambda^\bullet \dot{R}_n(\lambda^* + \varepsilon_n)$, and the second equality follows from Taylor expanding \dot{R}_n around λ^* . As a result,

$$\mathbb{L}^\bullet(\lambda^\bullet) = \frac{1}{2} (\lambda^\bullet)^2 \text{tr}\{\dot{\mathcal{R}}(\lambda^*)^\otimes \ddot{\mathcal{L}}^{-1}\} - \lambda^\bullet (\text{tr}\{\mathcal{I}_{\ell\ell} \ddot{\mathcal{L}}^{-1} \ddot{\mathcal{R}}(\lambda^*) \ddot{\mathcal{L}}^{-1}\} + 2\mathcal{B}^\top \dot{\mathcal{R}}(\lambda^*) - \text{tr}\{\mathcal{I}_{\ell r}(\lambda^*) \ddot{\mathcal{L}}^{-1}\} - \mathcal{E}).$$

When $\lambda^\bullet = 1$, then the expression reduces to $\mathbb{L}(\lambda^*) + \mathcal{E}$, proving the first part of the theorem.

Now, we optimize λ^\bullet . Define

$$A_1 \stackrel{\text{def}}{=} \text{tr}\{\mathcal{I}_{\ell\ell} \ddot{\mathcal{L}}^{-1} \ddot{\mathcal{R}}(\lambda^*) \ddot{\mathcal{L}}^{-1}\} + 2\mathcal{B}^\top \dot{\mathcal{R}}(\lambda^*) - \text{tr}\{\mathcal{I}_{\ell r}(\lambda^*) \ddot{\mathcal{L}}^{-1}\}, \quad (148)$$

$$A_2 \stackrel{\text{def}}{=} \text{tr}\{\dot{\mathcal{R}}(\lambda^*)^\otimes \ddot{\mathcal{L}}^{-1}\}. \quad (149)$$

Note that

$$\mathbb{L}^\bullet(1) = \frac{1}{2} A_2 - (A_1 - \mathcal{E}). \quad (150)$$

Now optimizing $\mathbb{L}^\bullet(\lambda^\bullet)$ with respect to λ^\bullet yields

$$\lambda^{\bullet*} = \frac{A_1 - \mathcal{E}}{A_2}, \quad \mathbb{L}^\bullet(\lambda^{\bullet*}) = \frac{-(A_1 - \mathcal{E})^2}{2A_2}. \quad (151)$$

Subtracting, we have

$$\mathbb{L}^\bullet(\lambda^{\bullet*}) - \mathbb{L}^\bullet(1) = -\frac{(A_1 - \mathcal{E} - A_2)^2}{2A_2}. \quad (152)$$

If $R_n(\lambda)$ is linear in λ , the linear and quadratic terms of the oracle relative risk must be balanced, so $A_1 = A_2 = -2\mathbb{L}(\lambda^*)$; as a special case, Corollary 1 applies when λ is a scalar. The second part of the theorem follows from algebra. \square

E Proof of Theorem 4

Proof. The theorem essentially follows by Taylor expanding f around θ_∞ . First, we have

$$\hat{\theta}_n^0 - \theta_\infty = -\ddot{L}_n^{-1} \dot{L}_n + O_p(n^{-1}). \quad (153)$$

Since f is differentiable,

$$\varepsilon_n = \hat{\lambda}_n - \lambda^* = -\dot{f} \ddot{L}_n^{-1} \dot{L}_n + O_p(n^{-1}), \quad (154)$$

where $\dot{f} \in \mathbb{R}^{b \times d}$ is the Jacobian of f . Plugging into the definition of \mathcal{E} (19), we get

$$\mathcal{E} = \lim_{n \rightarrow \infty} n^2 \mathbb{E}[\text{tr}\{\dot{L}_n (\nabla \dot{R}_n(\lambda^*) (-\dot{f} \ddot{L}_n^{-1} \dot{L}_n))^\top \ddot{\mathcal{L}}^{-1}\} + O_p(n^{-\frac{5}{2}})] \quad (155)$$

$$= \lim_{n \rightarrow \infty} -n^2 \mathbb{E}[\text{tr}\{\dot{L}_n^\otimes \ddot{L}_n^{-1} (\nabla \dot{R}_n(\lambda^*) \dot{f})^\top \ddot{\mathcal{L}}^{-1}\} + O_p(n^{-\frac{5}{2}})]. \quad (156)$$

Taking limits and observing that $\mathcal{I}_{\ell\ell}$ and $\ddot{\mathcal{L}}$ are symmetric completes the proof. \square