Policy Learning in High Dimensional Settings

Eddie (Yidi) Wu

Abstract

This paper studies policy learning in high-dimensional settings and documents a "double ascent" phenomenon in welfare, analogous to the double descent phenomenon in machine learning. Leveraging the equivalence between additive welfare maximization and weighted classification risk minimization, we analyze the performance of linear treatment assignment rules estimated via gradient descent as model complexity increases. We show via simulations that out-of-sample welfare is non-monotonic in model complexity, with a decline near the interpolation threshold followed by improvement in the overparameterized regime, demonstrating that more complex policies can potentially outperform parsimonious ones. Then, we derive weighted versions of the Vapnik–Chervonenkis generalization bound and margin-based linear model generalization bound tailored to policy learning. These results provide a general perspective for understanding how overparameterization can lead to welfare gains in high dimensional settings. Further simulations examine the robustness of the double ascent phenomenon under model misspecification and when using sieve estimators. Our findings suggest that overparameterization can be beneficial for welfare, particularly when rich covariate information is available, highlighting the potential value of flexible, high-dimensional policy rules.

1 Introduction

Policy learning is the process of using experimental or observational data to construct a personalized treatment assignment rule to maximize desirable objectives. More formally, for binary treatment, it learns a function $f : \mathcal{X} \mapsto \{+1, -1\}$ from a dataset $\mathcal{D}$ which maps individual features to treatment assignment. This paper draws upon the connection between welfare maximization in policy learning and weighted risk minimization in classification to study policy learning in high dimensional setting, which is defined as the regime where the number of features in the data is greater than the sample size.

By considering the class of linear classifiers with loss function weighted by observed outcome, we provide evidence for cases where population welfare derived from the optimal policy increases with model complexity in high dimensional settings, and under some data generation assumptions, overparameterized models can attain higher population welfare than underparameterized models. This phenomenon, which we term "double ascent", is analogous to double descent which is well-documented and widely studied for classification and regression [Belkin et al., 2019] [Belkin et al., 2020] [Hastie et al., 2022]. In the underparameterized regime, fixing sample size, generalization risk follows a U-shaped curve against model complexity as explained by overfitting and the classical bias-variance tradeoff. When model complexity grows beyond the point where it interpolates training data i.e. fits training data with zero error, generalization risk begins decreasing again with model complexity.

Prior work have advanced our theoretical understanding of this behavior by identifying specific data generating processes (DGP) and model classes under which it occurs and mechanisms that account for it. In this paper, following the perspective of Lee and Cherkassky [2024] who proposes the idea of analyzing double descent through the lens of the Vapnik-Chervonenkis (VC) theoretic framework, we derive weighted version of the VC generalization error bounds to understand welfare double ascent when using overparameterized linear treatment assignment rules. These results suggest that welfare improvement in the overparameterized regime can be explained by implicit control of model complexity even as the feature dimension increases.

Through additional simulations, we study settings in which the true DGP is low-dimensional while the treatment assignment rule uses sieve-based feature expansions, including polynomial, spline, random ReLU and random Fourier features. We evaluate how welfare behaves as sieve becomes more complex and find that the sieve estimators exhibit a double ascent pattern once sieve features become sufficiently expressive to interpolate the training data. Furthermore, to better understand the performance of linear treatment assignment rules with implicit bias, we compare linear classifiers to shrinkage-based estimators such as LASSO and ridge which are commonly used in high dimensional settings. Lastly, we benchmark the performance of learning treatment assignment rule directly via weighted classification against the approach of learning conditional average treatment effect (CATE) and then implementing treatment assignment policy via plug-in CATE.

In terms of significance for practitioners, one key takeaway is that, under certain DGPs and classes of treatment assignment rules, practitioners may benefit from using as much information as possible about the units in the sample. Leveraging richer feature representations can lead to higher welfare relative to more parsimonious specifications. As shown

in this study, one such example is a linear classifier trained via gradient descent. Other examples include function classes that have been shown in previous work to exhibit "double descent" in generalization error. These findings are especially relevant in today's world where big data is readily available. That said, concerns about interpretability and real world constraints such as policymakers' preference for simple decision rule involving few variables and the financial cost of collecting individual information may pose potential hurdles.

1.1 Related work

A large body of work studies the policy learning problem. Manski [2004] introduces the conditional empirical success rule which assigns persons to treatments based on the best observed experiment outcomes within subgroups defined by covariates, and derives a closed-form bound on the maximum regret of such rules. Kitagawa and Tetenov [2018] study the empirical welfare maximization (EWM) method which maximizes the sample analog of average social welfare over a class of candidate treatment policies and show that the average welfare attained by EWM rule converges to the maximum attainable welfare at a minimax optimal rate. Athey and Wager [2021] take a semiparametric approach by estimating outcome and propensity score using cross-fitting and constructing doubly robust scores to maximize welfare over a policy class. Their method attains minimax-optimal regret bounds and $O(n^{-1/2})$ convergence rate in welfare. While these work primarily focus on asymptotics in the sample size n, we study the setting where n is fixed but the feature dimension increases, and investigate the welfare performance of linear treatment assignment rule under known propensity score.

A line of work formulates learning optimal treatment assignment rule as a weighted risk minimization problem. Zadrozny [2003] first proposes cost-sensitive learning which minimizes example dependent misclassification costs instead of error rate to extract decision policies from data. Zhao et al. [2012] show that estimating treatment rule is equivalent to a classification problem in which units are weighted by clinical outcome and propose an outcome weighted learning approach based on support vector machine. Kitagawa et al. [2023] cast causal policy learning as a weighted classification problem, study surrogate loss consistency under misspecification and develop robust hinge loss-based procedures for monotone classification and policy learning. Motivated by this perspective, we study policy learning in high dimensions via weighted classification with linear decision rules and analyze welfare performance through the lens of misclassification cost.

An extensive literature analyze the generalization behavoir of classifiers and regressors, particularly when models exhibit vanishing error on the training data [Liang and Recht, 2023] [Montanari et al., 2023]. Belkin et al. [2019] and Nakkiran et al. [2021a] document evidence for the existence of double descent across a broad range of models and datasets, especially in modern deep learning tasks. To gain theoretical insights into double descent, Belkin et al. [2020] provides a concrete formal analysis of double descent in tractable settings using Gaussian and Fourier series models with least squares and least norm predictors. For regression problems, Bartlett et al. [2020] characterize linear regression in terms of effective rank of the data covariance and show that the minimum norm interpolator has near optimal prediction accuracy. Hastie et al. [2022] further consider least squares regression to demonstrate double descent of prediction risk and propose that it can be explained by the

reduction in variance as feature dimension increases. For classification problems, Deng et al. [2020] and Kini and Thrampoulidis [2020] study the double descent behavior of prediction error of linear classifiers trained with logistic loss and square loss respectively for data generated from logistic model and Gaussian mixture model. Spiess et al. [2023] also provide a model-averaging interpretation for double descent in high dimensional linear regression.

Other attempts to explain double descent in high dimensions include Gu et al. [2024] who mechanically study the learned feature space and find that although overparameterized models interpolate noisy data, they learn latent representation that encode the true data structure. Cherkassky and Lee [2024] and Lee and Cherkassky [2024] put forth the notion of understanding double descent more generally using the VC theoretical framework. The second descent corresponds to minimizing the VC dimension of the set of models with near zero training error and this is supported by empirical evidence from learning methods such as SVM. Although this is not a precise explanation since VC bound is a uniform risk upper bound for an entire class of functions rather than characterization of a specific estimator, it offers an appealing perspective nevertheless and our analysis builds on this line of thought. Additionally, a crucial class of work studies the role of implicit bias in optimization algorithms. For instance, gradient descent can bias models towards global minima with good generalization performance [Neyshabur et al., 2017] [Soudry et al., 2018]. Based on these insights, we adopt the VC dimension perspective of Lee and Cherkassky [2024] and interpret the performance gain of overparameterization as arising from complexity control induced by the implicit bias of the optimization algorithm.

1.2 Outline

In Section 2, we describe our simulation set-up and present evidence of welfare double ascent when using linear classifier trained via gradient descent and that welfare performance in the overparameterized regime dominates the underparameterized regime under certain DGPs. These DGPs can be highly relevant to many real-world settings.

In Section 3, we develop theoretical results based on VC generalization bounds for weighted population classification risk, analyze the behavior of linear classifier trained with weighted logistic loss in high dimensional settings, and discuss potential mechanisms underlying welfare gains in the overparameterized regime.

In Section 4, we provide further simulation evidence comparing the welfare performance of linear classifier with shrinkage estimators including ridge and LASSO. Then, we investigate double ascent under model misspecification where the true DGP is low dimensional and nonlinear, but we fit linear sieve estimators. Additionally, we compare the welfare performance of plug-in policies derived from CATE estimated by T-learner, with policies learned directly via weighted classification, under both correctly specified and misspecified models.

In Section 5, we conclude with a discussion on the advantages of weighted classification for policy learning in high dimensional settings, along with open questions and directions for future research.

In Appendix A, we provide additional simulation evidence for double ascent in cases where the true DGP varies with feature dimension. We also illustrate cases where overparameterization using sieve estimators does not yield performance improvements and give potential explanations for such behavior.

2 Welfare Double Ascent

2.1 Set-up and notations

In binary classification, let $\mathbf{x_i} \in \mathcal{X}$ be the vector of features, $t_i \in \{+1, -1\}$ be the target, and ω_i be the weight or classification cost. ω_i can be a function of $\mathbf{x_i}$ or t_i. Suppose $(\mathbf{x_i}, t_i) \sim P$, and the classification rule is $sign(f(x_i))$ for some $f : \mathcal{X} \mapsto \mathbb{R}$, then the population weighted classification risk as a function of f is expressed as:

$$R^{\omega}(f) = E_P[\omega_i \mathbf{1}\{t_i sign(f(\mathbf{x_i})) < 0\}]$$

This is analogous to the usual classification risk except that each unit is weighted by ω_i rather than equally. Fixing a sample of $\{\mathbf{x_i}, t_i\}_{i=1,..n}$, and suppose ω_i can be computed from the sample, the empirical weighted classification risk is:

$$\hat{R^{\omega}}(f) = \frac{1}{n}\sum_{i=1}^{m} \omega_i \mathbf{1}\{t_i sign(f(\mathbf{x_i})) < 0\}$$

Now, consider the problem of maximizing additive welfare objective under the usual Neyman-Rubin causal framework, let $\mathbf{x_i} \in \mathcal{X}$ be the vector of features as in the classification problem, $D_i \in \{+1, -1\}$ be the treatment status, $y_i(+1)$ and $y_i(-1)$ be the treated and untreated potential outcomes respectively, and $y_i = y_i(-1) + \frac{D_i+1}{2}(y_i(+1) - y_i(-1))$ be the observed outcome. We let $e(x_i) = Pr(D_i = 1|\mathbf{x_i})$ be the propensity score, and suppose $(y_i(+1), y_i(-1), D_i, \mathbf{x_i}) \sim Q$. We denote the treatment assignment rule as f where units with $f(\mathbf{x_i} \geq 0)$ receive treatment. Then the population welfare objective as a function of f can be expressed as:

$$W(f) = E_Q[y_i(+1)\mathbf{1}\{f(\mathbf{x_i}) \geq 0\} + y_i(-1)\mathbf{1}\{f(\mathbf{x_i}) < 0\}]$$

As shown in Kitagawa et al. [2023], the following proposition establishes the equivalence between weighted risk minimization and additive welfare maximization.

Proposition 2.1 (Risk minimization and welfare maximization [Kitagawa et al., 2023]) *Assuming $e(x)$ and Q satisfy the following properties:*

- *Overlap: propensity score $0 < e_{\min} \leq e(x) \leq e_{\max} < 1$ for all $x \in \mathcal{X}$*
- *Unconfoundedness: $y_i(+1), y_i(-1) \perp D_i|\mathbf{x_i}$*

Then the maximization of additive welfare criterion $W(f)$ is equivalent to the minimization of weighted classification risk

$$E_Q[\omega_i \mathbf{1}\{t_i sign(f(\mathbf{x_i})) < 0\}]$$

where $t_i = sign(y_i)D_i$ and $\omega_i = \frac{|y_i|}{D_i e(\mathbf{x_i}) + (1-D_i)/2}$.

Proof: Express welfare objective in terms of observed outcome:

$$
\begin{aligned}
W(f) &= E_Q[y_i(+1)\mathbf{1}\{f(\mathbf{x_i}) \geq 0\} + y_i(-1)\mathbf{1}\{f(\mathbf{x_i}) < 0\}] \\
&= E_Q[\frac{y_i\mathbf{1}\{D_i = +1\}\mathbf{1}\{f(\mathbf{x_i}) \geq 0\}}{e(\mathbf{x_i})} + \frac{y_i\mathbf{1}\{D_i = -1\}\mathbf{1}\{f(\mathbf{x_i}) < 0\}}{1 - e(\mathbf{x_i})}] \\
&= E_Q[\frac{y_i\mathbf{1}\{D_i = +1\}\mathbf{1}\{sign(f(\mathbf{x_i})) = D_i\}}{e(\mathbf{x_i})} + \frac{y_i\mathbf{1}\{D_i = -1\}\mathbf{1}\{sign(f(\mathbf{x_i})) = D_i\}}{1 - e(\mathbf{x_i})}] \\
&= E_Q[\Big(\frac{y_i\mathbf{1}\{D_i = +1\}}{D_i e(\mathbf{x_i}) + (1 - D_i)/2} + \frac{y_i\mathbf{1}\{D_i = -1\}}{D_i e(\mathbf{x_i}) + (1 - D_i)/2}\Big)\mathbf{1}\{sign(f(\mathbf{x_i})) = D_i\}] \\
&= E_Q[\frac{y_i}{D_i e(\mathbf{x_i}) + (1 - D_i)/2} \cdot \mathbf{1}\{sign(f(\mathbf{x_i})) = D_i\}] \\
&= E_Q[\frac{y_i}{D_i e(\mathbf{x_i}) + (1 - D_i)/2} - \mathbf{1}(D_i sign(f(\mathbf{x_i})) < 0)\frac{y_i}{D_i e(\mathbf{x_i}) + (1 - D_i)/2}] \\
&= E_Q[\max\Big\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1 - D_i)/2}\Big\}] \\
&\quad - E_Q[\frac{|y_i|}{D_i e(\mathbf{x_i}) + (1 - D_i)/2}\mathbf{1}\{sign(y_i) \cdot D_i \cdot sign(f(\mathbf{x_i})) < 0\}] \qquad (1)
\end{aligned}
$$

The second last equality holds because when $D_i \neq sign(f(\mathbf{x_i}))$, the product $D_i sign(f(\mathbf{x_i}))$ is less than 0 and the terms inside the expectation sum to 0. The last line is equivalent to the second last line if one considers the cases of positive and negative y_i separately. □

The purpose of expressing welfare objective as shown in the last line of the proof is to ensure that the weight ω_i is always positive. In the last line, since f only affects the second term, choosing f to minimize the second term is equivalent to maximizing $W(f)$.

To see the intuition behind the equivalence, let's suppose that y_i is always positive for simplicity. Then, the classification problem becomes minimizing the weighted version of $\mathbf{1}\{D_i \cdot sign(f(\mathbf{x_i})) < 0\}$, i.e. matching the sign of D_i to the sign of $f(\mathbf{x_i})$, with weights given by the inverse propensity-weighted observed outcome. If a unit has large observed outcome, then it is valuable for the model to match the treatment status D_i of the unit because mismatching $sign(f(\mathbf{x_i}))$ and D_i is costly. Misclassifying a unit with large observed outcome, i.e. treatment rule assigning the opposite treatment, incurs large cost in terms of welfare.

Another way of looking at what f learns in the weighted classification problem is by writing population welfare in terms of conditional average treatment effect $\tau(\mathbf{x_i})$:

$$W(f) = E_Q[y_i(-1)] + E_Q[\mathbf{1}\{sign(f(\mathbf{x_i})) \geq 0\}\tau(\mathbf{x_i})]$$

We can see that welfare is maximized when the policy assigns treatment to units with positive $\tau(\mathbf{x_i})$. This means that the welfare maximizing assignment rule f^* is attained when $sign(f^*(\mathbf{x_i})) = sign(\tau(\mathbf{x_i}))$. The weighted classification problem is essentially learning CATE but thresholding at 0, i.e. the classification decision boundary is also the boundary that separates positive and negative CATE.

Finally, it is notable that in the case of linearly separable treatment status labels and f being in a class of linear classifiers, doing welfare maximization by minimizing the empirical weighted loss $\hat{R}^{\omega}(f)$ might not appear meaningful because any hyperplane that separates

the treatment labels attains zero empirical weighted loss and the solution is not unique. That said, it is important to note that with separability, generalization depends on the inductive bias e.g. the choice surrogate loss, optimization algorithms and so on. For instance, replacing the discontinuous 0-1 loss with the convex logistic loss and using gradient descent (GD) for optimization leads to the maximum margin support vector classifier which has good generalization properties. We expound on this in more details in the latter sections of the paper. As such, maximizing welfare by minimizing empirical classification risk remains meaningful even under linearly separable labels.

2.2 Data generation process

We illustrate "double ascent" in additive welfare using the following RCT data generation set-up. n units $\{\mathbf{x_i}, D_i, y_i(+1), y_i(-1)\}_{i=1}^n$ are randomly drawn from:

$$\mathbf{x_i} \sim N_d(0, \sigma_x^2 I_d) \tag{2}$$

$$y_i(+1) = \mathbf{x_i}^T \beta_1 + \epsilon_{1,i} \tag{3}$$

$$y_i(-1) = \mathbf{x_i}^T \beta_0 + \epsilon_{0,i} \tag{4}$$

$$D_i \sim symmetric\ Bernoulli(q) \tag{5}$$

β_1, β_0, σ and p are pre-specified parameters. σ_x and p are assumed known by researcher while β_1 and β_0 are unknown. Under this set-up, CATE is $\mathbf{x}_i^T(\beta_1 - \beta_0)$ and in the absence of constraints on treatment allocation, the welfare maximizing treatment assignment rule, which we term the oracle rule, is $sign(\mathbf{x}_i^T(\beta_1 - \beta_0))$.

It is worth noting that the optimal policy in this set-up coincides with the risk minimizing decision boundary in the generative logistic model which is widely studied in binary classification problems. Let $g(z) := (1 + e^{-z})^{-1}$ be the sigmoid function, the generative logistic model generates data from

$$t_i = \begin{cases} +1 & w.p.\ g(\mathbf{x_i}^T \alpha) \\ -1 & w.p.\ 1 - g(\mathbf{x_i}^T \alpha) \end{cases}$$

In this classification problem, a random sample of $\{\mathbf{x_i}, t_i\}$ is used to learn the risk minimizing classifier, whereas in the policy learning problem, a random sample of features, treatment status, and observed outcome is used to learn the welfare maximizing assignment rule. If $\alpha = (\beta_1 - \beta_0)$, the risk minimizing classifier is the same as the optimal assignment rule.

Deng et al. [2020] characterizes double descent of classification risk when the true DGP is the logistic model and binary classifier is used with gradient descent. Despite sharing the same optimal decision boundary, our set-up differs in two ways. First, Deng et al. [2020] studies the usual binary classification problem while policy learning requires weighted classification. Second, binary classifications learns from label of $\{+1, -1\}$ directly while policy learning uses observed outcome and treatment status.

2.3 Learning algorithm

In this simulation experiment, we assume correct model specification and focus on linear policy rules i.e. treatment assignment is given by $sign(\mathbf{x_i}^T\theta)$. Instead of minimizing the weighted empirical 0-1 classification risk directly which is non-convex and computationally intractable, we minimze the empirical surrogate risk:

$$\hat{R}^{\omega}_{\phi}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}\omega_i\phi(t_i f(\mathbf{x_i})) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\phi(t_i\mathbf{x_i}^T\theta)$$

where we choose $\phi(z) = \log(1+e^{-z})$ which is convex and infinitely differentiable everywhere. Mean squared loss $\phi(z) = (1-z)^2$ is another viable choice of surrogate as shown by Kini and Thrampoulidis [2020] and Liang and Recht [2023].

However, we note that according to Kitagawa et al. [2023], in our study under high dimensional setting, the class of classifiers $\mathcal{F}$ is constrained and R_ϕ-misspecified in most cases, which means $\inf_{f\in\mathcal{F}} R_\phi(f) > R_\phi(f^*)$ where f^* is the minimizer of R_ϕ over all measurable functions in the terminology of Kitagawa et al. [2023], i.e. minimizing surrogate loss over $\mathcal{F}$ does not lead to the global minimizer of surrogate loss. As such, their Proposition 2.1 implies that using logistic loss might not minimize the original 0-1 loss. Moreover, consider the case when features in the true DGP are omitted from the set of features in the linear model, this is equivalent to restricting some coefficients to 0. This is a form of $\mathcal{G}$-constrained classification and according to Proposition 2.2 in Kitagawa et al. [2023], the constrained function class $\mathcal{F}_\mathcal{G}$ is neither R- or R_ϕ- correctly specified. Under R-misspecification, Kitagawa et al. [2023] Theorem 3.2 and Corollary 3.7 show that hinge loss is the only surrogate loss function, hence excluding logistic and squared loss, that ensures the consistency of surrogate loss minimizer to the classification risk minimizer over $\mathcal{F}_\mathcal{G}$. However, because linear model imposes further functional form restriction on $\mathcal{F}_\mathcal{G}$, even hinge loss is not guaranteed to be consistent in this case.

Therefore, using logistic loss does not guarantee that minimal surrogate risk over the class of linear functions equals the minimal classification risk over the class of linear functions. The main reason for choosing logistic loss is due to its attractive properties in high dimensional settings with gradient descent, and its computational tractability. We expound more on this in later sections.

We optimize the linear policy rule using gradient descent (GD). Starting from a randomly initialized value of $\alpha_{(0)}$, for each iteration $k > 0$, we update the parameter estimate by:

$$\alpha_{(k+1)} = \alpha_{(k)} - \eta_k \nabla \hat{R}^{\omega}_{\phi}(\alpha_{(k)})$$

We iterate this process for a fixed number of steps to mitigate the effect of early stopping in GD as feature dimension changes, and set the number to be very large so that parameter estimate is close to convergence. We also try iterating until the stopping criteria, such as the change in loss and the gradient magnitude, exceed some specified threshold. This does not alter the conclusions of our simulations.

To investigate the effect of overparameterization, we fix the sample size n throughout and consider two different approaches of varying feature dimensionality popularly used in

the literature. The first approach sets the true DGP to have feature dimension $\mathbf{x}_i \in \mathbb{R}^d$, and generate random neural features from $\mathbf{x_i}$ by:

$$\mathbf{z_i} = \sigma(\mathbf{x_i}^T W)$$

where $\sigma(\cdot)$ is an element-wise nonlinear activation function, $W \in \mathbb{R}^{d \times \zeta}$ is matrix of random weights and $\mathbf{z}_i$ is in $\mathbb{R}^\zeta$. For each integer value of feature dimension $p \in [1, \zeta]$, we take the first p elements of $\mathbf{z}_i$ and fit a linear model to estimate an empirical loss-minimizing parameter vector of length p. This means that for each overparameterizing ratio $\frac{p}{n}$, we are using the same information in $\mathbf{x_i}$ but artifically increasing the dimensionality of input features to investigate overparameterization. Each element of W is sampled independently from $N(0, 1/d)$ to ensure that each $\mathbf{x}_i^T W_k$, where W_k is column k of W, has approximately zero mean and unit variance and the resulting random ReLU features are of similar scale.

The second approach considers omitted variables. In the true DGP, $\mathbf{x_i} \in \mathbb{R}^\zeta$ is sampled from multivariate standard normal distribution. For each integer value of $p \in [1, \zeta]$, we take a subset of features in $\mathbf{x_i}$ with size p to fit a linear model. We assume each element $\alpha_k := (\beta_{1,k} - \beta_{0,k})$ in the true parameter vector α has similar signal strength, so we can take the first p features in $\mathbf{x_i}$ for each p WLOG. We also normalize the subset feature vector by $\sqrt{p}$ for each dimension p to maintain similar signal strength across dimensions, so as to isolate the effect of overparameterization. Except for the model with the full set of features of $\mathbf{x_i}$, all the other models are misspecified. We note that another commonly seen method of defining α is such that signal strength decays at a polynomial rate as the number of features increases and this does not alter our conclusions.

2.4 Simulation results

In our simulations, we set the training data size n to be 150 and the test data size for computing generalization risk or welfare to be 5000. In the random ReLU feature model, we set the dimension of $\mathbf{x}_i$ as 50 and the maximum dimension of random ReLU features ζ to be 300, so the highest overparameterization ratio $p/n = 2$. In the omitted variable model, we set the dimension of features in the true DGP to be 300. We draw β_0 from $N(-1, 4I)$ and β_1 from $N(1, 4I)$, and set the RCT probability of being treated q to be 0.5. We draw $\mathbf{x_i}$ from $N(0, I)$, and $\epsilon_{0,i}$ and $\epsilon_{1,i}$ from $N(0, 1)$.

During training, we use gradient descent with logistic loss and a fixed learning rate $\eta_k = 0.01$ for all k, minimizing the empirical surrogate loss to estimate the parameter vector $\hat{\alpha}$. Learning rate is chosen by monitoring training loss over GD iterations to ensure that loss decreases monotonically to a minimum. After obtaining $\hat{\alpha}$, we compute in-sample welfare by $\frac{1}{n}\sum_{i=1}^n (y_i(+1)\mathbf{1}\{f(\mathbf{x_i}) \geq 0\} + y_i(-1)\mathbf{1}\{f(\mathbf{x_i}) < 0\})$ since we have the values of potential outcomes. Out-of-sample population welfare is computed using the test set in the same way. We average the results over 100 simulation runs and monitor

Figure 1 plots welfare against the number of random ReLU features. Interpolation threshold is the p/n ratio beyond which linear classifier interpolates the training data in terms of weighted classification risk and training welfare attains a stable value. In the underparameterized regime, test welfare improves rapidly as p increases, reflecting greater ability to approximate the decision boundary as more random ReLU features are introduced. Near the interpolation threshold, training welfare saturates while test welfare deteriorates.

In the overparameterized regime, test welfare shows a second ascent and continues to improve as the number of features increases. The blue dotted line labelled "Using original X's" is the test welfare obtained using the same sample but fitting the original 50 features in $\mathbf{x}_i$. It is not surprising that the test welfare of random ReLU features approaches the original $\mathbf{x}_i$'s welfare as p increases because the random ReLU features approximate the true linear DGP through sufficiently rich feature expansions. Despite being highly overparameterized, the model shows strong generalization performance.

This experiment is relevant to real-world settings in the sense that random ReLU features act as noisy feature expansion of the covariates in the true DGP. It could be that the underlying true covariates are not directly observed and random ReLU features can be viewed as a large collection of many observable weak nonlinear transformations of the underlying signals. Increasing such features enriches the representation of the true DGP even when fixing sampling size.

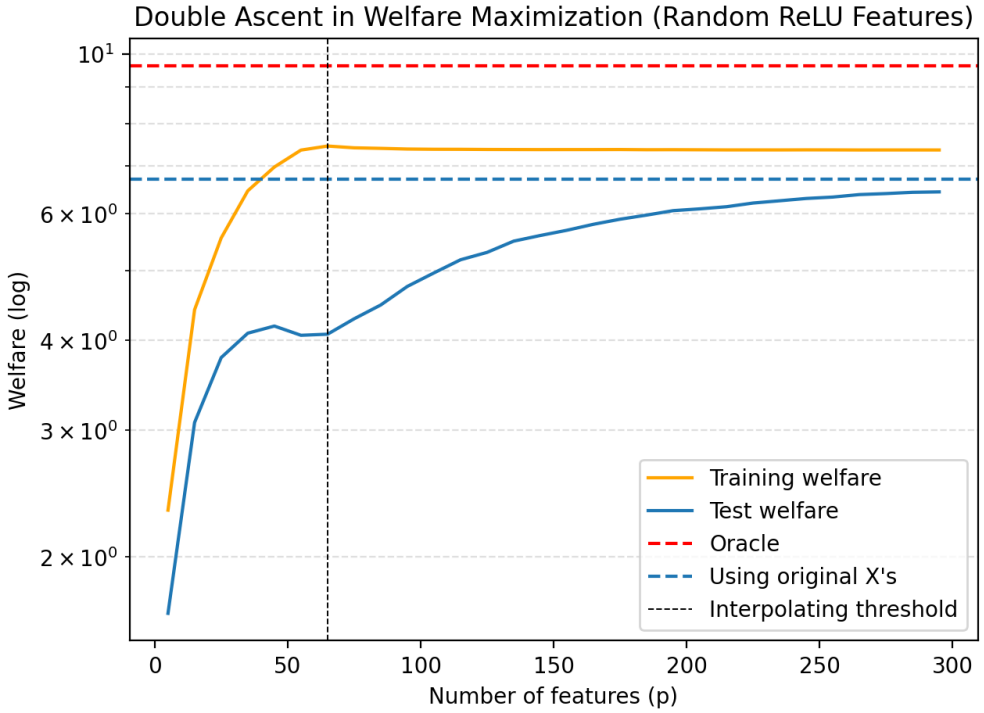

Figure 1: Welfare against feature dimension using random ReLU features. Training data size n is 150 and the number of features grows from 5 to 300. Training welfare attains a stable value at the interpolation threshold while test welfare experiences a second ascent beyond interpolation. The blue dashed line represents welfare performance using the original features in the DGP. Increasing the number of random ReLU features enhances generalization.

Figure 2 plots welfare against the number of features included in the classifier in the omitted variables model. The blue dashed curve labelled "Second best oracle" illustrates the maximum welfare attainable when a subset of features of size p is included in the classifier, and it converges to the oracle welfare as the p approaches 300.

Test welfare increases rapidly in the underparameterized regime, falling towards the

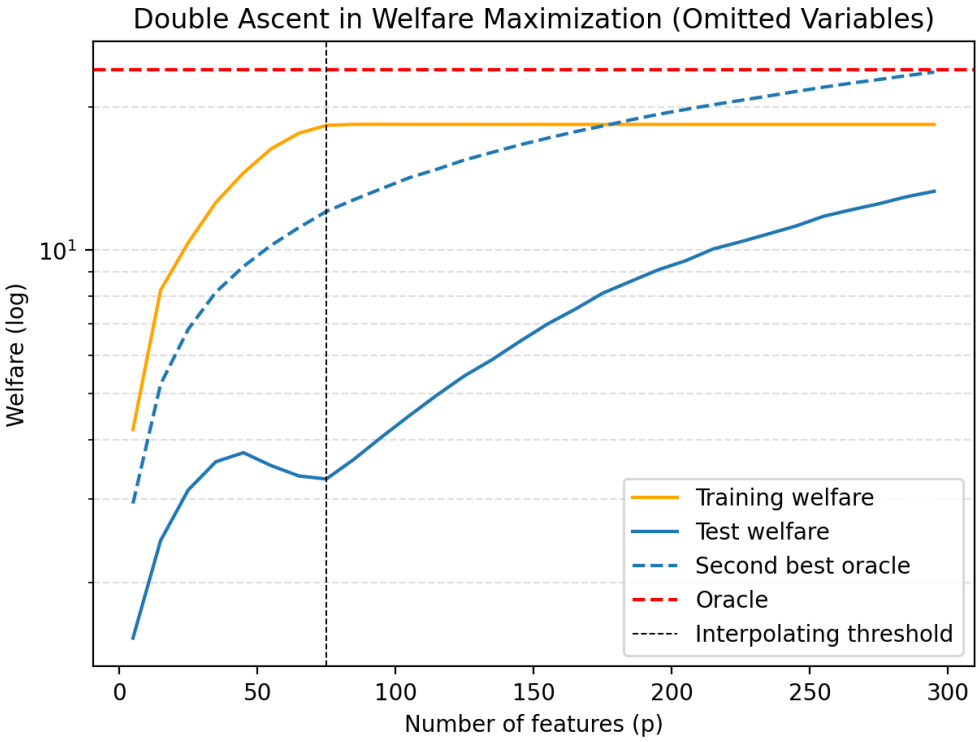

Figure 2: Welfare against feature dimension in the omitted variables model. Training data size n is 150 and the number of features included grows from 5 to 300. Training welfare attains a stable value at the interpolation threshold while test welfare experiences a second ascent beyond interpolation. The blue dashed curve represents the maximum welfare attainable using p number of features. Increasing the number of features included in the model enhances generalization.

interpolation threshold and increases gradually again in the overparameterized regime. There are two effects at play here. First, the gap between oracle welfare and second best oracle welfare is increasing in $P(sign(\mathbf{x}^T(\beta_1 - \beta_0)) \neq sign(\mathbf{x}^T(\beta_{1,p} - \beta_{0,p})))$, where $\beta_{1,p}$ is the vector containing the first p elements of β_1. In our DGP, it is not difficult to show analytically that this probability is strictly decreasing in p, i.e. the fewer features we omit, the higher the second best oracle welfare. As such, although we normalize $\mathbf{x}_i$ by $\sqrt{p}$ to maintain similar signal strength across p, including more features helps with generalization performance. The second source of improvement comes from overparameterization where expanding the feature space and implicit bias lead to improvement in generalization. We discuss this in more details in the next section.

The omitted variable model closely reflects many real-world settings, where the learner observes only a subset of true features that drive outcome in the true DGP. The experiment shows that augmenting the model with additional relevant features is beneficial to learning. Notably, in both set-ups, higher test welfare is attained in the overparameterized setting, suggesting potential benefits of using overparameterized models for policy learning. Also, test welfare is consistently lower than training welfare, mirroring how training risk is often lower than test risk in classification.

3 Weighted Generalization Bounds

In Section 2, we show that double descent phenomenon also appears for the weighted classification risk that arises in policy learning. In this section, we seek to understand potential mechanisms for double descent (or double ascent) using the VC theoretical framework.

To the best of our knowledge, Lee and Cherkassky [2024] are the first proponents of VC theoretical explanations of double descent phenomenon. Generalization bounds are useful here because they decompose population risk into an empirical fit term and a model complexity term which is especially relevant for studying overparameterization. Once the model class is sufficiently complex to fit the training data perfectly, the empirical risk term is essentially zero for many candidate classifiers, so the complexity term determines the upper bound on their generalization risks.

The VC theoretic framework suggests one possible explanation for double descent is that, among the interpolating classifiers, the effective complexity of more overparameterized models may be smaller than that of simpler models, leading to tighter generalization upper bound and better test performance. To elaborating on effective complexity, while it is commonly believed that VC dimension of a hypothesis class increases with the number of parameters, Lee and Cherkassky [2024] emphasize that there is a distinction between VC dimension of a hypothesis class and that of a particular trained classifier. In overparameterized regimes, the learning algorithm may favor an interpolating solution with relatively small effective complexity, hence potentially improving generalization performace. We follow Lee and Cherkassky [2024]'s line of thought and extend their arguments to weighted classification and policy learning.

The extension is not automatic because instead of studying the usual 0-1 loss, under unconfoundedness, policy objective is written as a weighted classification objective. Therefore, we start by establishing VC-type generalization bounds focusing on weighted population

risk, with reference to Mohri et al. [2018] who provide the original VC bounds.

We first assume that the weights, which take the form of inverse propensity weighted (IPW) observed outcome in policy learning $\frac{y_i}{D_i e(\mathbf{x_i})+(1-D_i)/2}$, are bounded. This assumption is satisfied if the observed outcome is bounded and the propensity score is bounded away from 0. The second condition is automatically satisfied by the design of RCT or by making the overlap assumption which requires propensity score to be bounded away from 0. The first condition of a bounded observed outcome arises naturally in many applications where policy outcomes are like rates, probabilities or ordinal measures. It is also consistent with common empirical practice such as normalization or winsorization to bound observed outcome. The following proposition shows how weighted population risk is controlled by empirical weighted risk and VC dimension:

Proposition 3.1 (VC-dimension generalization bound for weighted classification) *Let $\mathcal{H}$ be a class of functions taking values in $\{+1, -1\}$ with VC-dimension d. Assume that the classification weight ω is bounded between $0 \leq \omega \leq W$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size n, the following holds for all $f \in \mathcal{H}$:*

$$R^{\omega}(f) \leq \hat{R}^{\omega}(f) + W\left(\sqrt{\frac{2d\log\frac{en}{d}}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \tag{6}$$

$$= O\left(W\sqrt{\frac{\log\frac{n}{d}}{\frac{n}{d}}}\right) \tag{7}$$

where $R^{\omega}(f)$ and $\hat{R}^{\omega}(f)$ are defined in Section 2.1.

Proof: Please refer to Appendix B.1. □

The bounded weights assumption is convenient serving as a useful baseline. To make it less restrictive, the next proposition is analogous to Proposition 3.1 but it relaxes the assumption that the weights are bounded and instead assume that they are sub-Gaussian. This allows for more realistic tail behavior of outcome in many policy learning settings. For instance, sub-Gaussian observed outcome combined with RCT implies that the weights are sub-Gaussian in policy learning.

Proposition 3.2 (VC-dimension generalization bound with sub-Gaussian weights) *Let $\mathcal{H}$ be a class of functions taking values in $\{+1, -1\}$ with VC-dimension d. Assume that the classification weight ω is sub-Gaussian with constant K. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size n, the following holds for all $f \in \mathcal{H}$:*

$$R^{\omega}(f) \leq \hat{R}^{\omega}(f) + K\sqrt{\log\frac{2n}{\delta}}\left(2\sqrt{\frac{d\log\frac{en}{d}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}\right) \tag{8}$$

$$= O\left(\sqrt{\log n}\sqrt{\frac{\log\frac{n}{d}}{\frac{n}{d}}}\right) \tag{9}$$

Proof: Please refer to Appendix B.2. □

Comparing the two bounds, the main difference is that the second bound replaces the maximum weight W with a term on the order of $\sqrt{\log n}$ which can be interpreted as the additional cost of allowing weights to be unbounded but sub-gaussian. Proposition 3.2 shows that weighted population risk is bounded above by the sum of weighted empirical risk and a second term that depends on both sample size and VC-dimension of $\mathcal{H}$. Rewriting:

$$\sqrt{\frac{d \log \frac{en}{d}}{n}} = \sqrt{\frac{1 + \log(\frac{n}{d})}{\frac{n}{d}}}$$

this says that the complexity term is governed by the ratio n/d. In the domain where $\frac{n}{d} \geq 1$, this expression is strictly decreasing in $\frac{n}{d}$. As sample size n increases, the second term goes to zero and $\hat{R}^{\omega}(f)$ dominates the upper bound on weighted population risk.

The effect of VC dimension d on the bound is more nuanced. Suppose the training sample size n is fixed, increasing model complexity improves in-sample fit and thereby reducing $\hat{R}^{\omega}(f)$. At the same time, a richer hypothesis class has larger VC-dimension, which might decrease n/d and inflate the complexity penalty. In the underparameterized regime, this creates a tradeoff between empirical fit and complexity, mirroring the classical bias-variance tradeoff. Optimal generalization performance is therefore achieved at an intermediate level of model complexity.

In the overparameterized regime, the situation changes. When the hypothesis class is complex enough to interpolate the training data, the empirical weighted risk can be driven to zero among numerous candidate fitted models and the generalization bound is governed entirely by sample size and VC-dimension. The prevalent interpretations of VC theory would suggest that VC dimension of the hypothesis class typically grows with the number of parameters and VC bound becomes inapplicable in high dimension. However, as argued by Lee and Cherkassky [2024], if we distinguish between complexity of the hypothesis class and that of the fitted classifier selected by the learning algorithm, VC dimension can in fact be decreasing in model complexity, leading to better generalization upper bound.

Before applying this framework to our set-up of linear classifier optimized with GD on weighted logistic loss, we first translate the generalization bound on weighted classification risk into a corresponding bound for population welfare, expressed in terms of empirical welfare and the VC dimension of the hypothesis class, as presented in the next proposition.

Corollary 3.1 (VC-dimension generalization bound for welfare) *Let $\mathcal{H}$ be a class of functions taking values in {+1, -1} with VC-dimension d. Suppose that the population distribution is Q, and that the classification weight, which is defined as $\omega := \frac{|y_i|}{D_i e(\mathbf{x_i}) + (1-D_i)/2}$, is sub-Gaussian with constant K. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size n, the following holds for all $f \in \mathcal{H}$:*

$$W(f) \geq \hat{W}(f) - K\sqrt{\log \frac{2n}{\delta}} \left(2\sqrt{\frac{d \log \frac{en}{d}}{n}} + \sqrt{\frac{\log \frac{2}{\delta}}{n}} \right) + C_{Q,n} \tag{10}$$

where n is the size of the training sample, $C_{Q,n}$ is a constant that depends only the population distribution Q and sample size n and is independent of f, and empirical welfare as

a function of treatment assignment rule f is defined as $\hat{W}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i(+1)\mathbf{1}\{f(\mathbf{x_i}) \geq 0\} + y_i(-1)\mathbf{1}\{f(\mathbf{x_i}) < 0\})$.

Proof: Please refer to Appendix B.3. □

Corollary 3.1 shows that population welfare is bounded below by sample welfare, up to a distribution dependent constant term and a complexity penalty term that depends on sample size and VC dimension. Fixing the hypothesis class, maximizing empirical welfare raises the lower bound for population welfare. However, when using an increasingly complex hypothesis class, model can attain higher empirical welfare but at the cost of greater penalty from the second term. In the overparameterized regime, empirical welfare is maxed out when f fits the sample perfectly, and further improvement in welfare lower bound can be attained by reducing model complexity and hence VC-dimension.

3.1 Weighted linear classifier in high dimensional setting

In the overparameterized regime, since model can interpolate training data and attain zero training loss, the optimization problem admits multiple global minima, with most of these minima not generalizing well to new observations. However, it is documented that optimization algorithms such as gradient descent exhibit implicit bias i.e. inducing model to converge to particular solutions that generalize well. A canonical example is high dimensional linear regression, where minimizing squared loss using gradient descent yields the minimum-norm (ridgeless) interpolating solution which can have near optimal prediction accuracy under some conditions [Bartlett et al., 2020] [Hastie et al., 2022].

For binary classification, Soudry et al. [2018] show that when dataset is linearly separable and loss function is smooth strictly decreasing and non-negative, and its negative derivative has a tight exponential tail, gradient descent does not converge to a finite minimizer, but rather the norm of the parameter vector diverges while its direction converges to that of the max-margin SVM solution.

Building on this result, Deng et al. [2020] further characterize the double descent curve of classification error as a function of the overparameterization ratio for the max-margin SVM. In this paper, we extend these insights to the weighted loss setting and find that a linear classifier optimized using gradient descent on weighted logistic loss likewise converges in direction to the unweighted max-margin SVM solution.

To elaborate on this, in the underparameterized regime where the design matrix $X \in \mathbb{R}^{n\times d}$ has full column rank, the weighted logistic loss function is strictly convex in the parameter vector θ and admits a unique minimizer. Consequently, gradient descent converges to this finite unique solution. On the other hand, in the overparameterized regime, data becomes linearly separable and weighted empirical risk admits no finite minimizer. Along the gradient descent trajectory, logistic loss decreases towards 0 as $||\theta|| \to \infty$. Despite this divergence in norm, the direction of θ stabilizes and converges to that of the max-margin SVM solution.

Since classification and hence policy welfare only depends on the sign and not magnitude of $\mathbf{x}_i^T\theta$, the direction of θ is what matters for generalization. As a result, an overparameterized linear classifier attains approximately similar generalization performance as a max-margin SVM if GD is run sufficiently long.

The following proposition puts forth the above claim formally:

Proposition 3.3 (Convergence to SVM with Weighted Loss) *Consider a dataset* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *with* $y_i \in \{+1, -1\}$ *that is not measure zero, and assuming that the data is separable i.e.*

$$\exists \theta \ s.t. \ y_i \mathbf{x}_i^T \theta > 0, \quad \forall i$$

Let ω_i *be fixed sample weights s.t.*

$$0 < \omega_{min} \leq \omega_i \leq \omega_{max} < \infty, \quad \forall i$$

Assume that the surrogate loss function ϕ *and learning rate* η *satisfy the conditions in Theorem 3 of Soudry et al (2018). Let* $\theta(t)$ *be the parameter vector at iteration* t *of gradient descent, and consider running gradient descent on* $\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \omega_i \phi(y_i \mathbf{x_i}^T \theta)$*, then the following holds:*

$$\lim_{t \to \infty} \frac{\theta(t)}{||\theta(t)||} = \frac{\hat{\theta}}{||\hat{\theta}||} \tag{11}$$

where $\hat{\theta}$ *is the solution of the max-margin SVM fitted on* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Proof: Please refer to Appendix C. □

An intuition for the above proposition can be seen from the max-margin SVM problem formulation, which solves:

$$\min_{\theta} \quad \text{subject to} \quad y_i \mathbf{x}_i^T \theta > 1 \quad \text{for all } i.$$

Introducing weights in the logistic loss can be viewed heuristically as replicating observations in the training sample. Such reweighting does not alter which points are support vectors in the separable case, and hence does not change the resulting max-margin hyperplane.

We now apply the weighted VC bounds to analyze welfare double ascent in our simulations. As feature dimension p increases with the sample size fixed, and assuming no perfect collinearity among features, a phase transition occurs at or before $p \approx n$, where the data becomes linearly separable and linear classifier interpolates training data. Prior to the transition, the linear classifier solution corresponds to the finite weighted logistic loss minimizer which is also the maximum likelihood solution. After the interpolation happens, linear classifier converges to the max-margin SVM direction and such implicit bias induces a form of effective regularization i.e. among all interpolating classifiers, gradient descent selects the direction of the minimum norm separator.

While the VC dimension of a class of linear classifiers in $\mathbb{R}^k$ is $k + 1$, a sharper characterization exists for Δ-margin classifiers. Vapnik [2000] shows in Theorem 5.1 that, suppose $\mathbf{x}$ belongs to a sphere of radius r, then the set of Δ-margin separating hyperplanes (support vector classifier) has its VC-dimension d bounded by:

$$d \leq \min(\left[\frac{R^2}{\Delta^2}\right], k) + 1$$

Note that max-margin SVM minimizes $||\theta||^2$ subject to $y_i \mathbf{x}_i^T \theta \geq 1$, and the margin Δ is equal to $\frac{1}{||\theta||}$. As the feature dimension increases further in the overparameterized regime, fixing sample size,additional coordinates can be introduced without increasing the minimum norm of the classifier e.g. by assigning zero coefficients to the new features. As such, the optimal value of $||\theta||$ does not increase and the margin $\Delta = \frac{1}{||\theta||}$ is non-decreasing. Since the VC-dimension bound scales inversely with Δ^2, this implies that the effective complexity of the max margin classifier does not increase and may in fact decrease in feature dimensions.

Therefore, in the overparameterized regime, weighted empirical risk is zero and we can write the upper bound on weighted population risk in Proposition 3.2 in terms of θ (normalizing $r = 1$) for the case of linear classifier trained with GD:

$$R^{\omega}(f) \leq K\sqrt{\log\frac{2n}{\delta}}\left(2\sqrt{\frac{d\log\frac{en}{d}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}\right) \tag{12}$$

$$\leq K\sqrt{\log\frac{2n}{\delta}}\left(2\sqrt{\frac{||\theta||^2\log\frac{en}{||\theta||^2}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}\right) \tag{13}$$

in the domain where $\frac{n}{||\theta||^2} \geq 1$. This upper bound on weighted population risk shrinks as we increase the dimension of $\mathbf{x}$ (normalizing it to have a constant radius r). It can be interpreted as a form of implicit regularization as gradient descent biases solution towards one with the smallest norm among all interpolating solutions. Connecting it to Corollary 3.1, the lower bound on population welfare increases with feature dimension in the overparameterized regime.

Figure 3 reports the training weighted logistic loss and the classification risks (i.e. misclassification rate of t_i using $sign(\mathbf{x}_i^T\theta)$) for training and test at the end of GD versus feature dimension for both random ReLU feature model and omitted variables model. We observe that both loss and classification risks drop to zero during training and flattens at zero as the number of features increases, indicating that the linear classifier has interpolated training data. Additionally, we check that the cosine distance between the direction of the interpolating linear classifier and that of the max-margin SVM is close to 1. These provide evidence for the transition of linear classifier from the maximum likelihood solution to the max-margin SVM solution. Test classification risk in Figure 3 exhibits a noticeable double descent, mirroring the double ascent in welfare as shown in Figures 1 and 2.

Figure 4 illustrates the $L2$ norm of $\hat{\theta}$ obtained by fitting a hard-margin SVM on $\{(t_i, \mathbf{x}_i)\}_{i=1}^n$ directly. $||\theta||$ decreases monotonically in the number of features for both data models, implying that margin increases with dimensionality. This is consistent with the idea that larger margin enables the VC bound on weighted population risk to become smaller in the overparameterized regime.

However, it is important to emphasize that VC generalization bound constitutes only a uniform upper bound on the weighted population risk over a class of classifiers. As such, it characterizes an upper envelope on risk rather than the realized population risk of a specific estimator. While the bound itself may exhibit a double descent pattern as a function of the feature dimension, this does not imply that the risk of any particular fitted classifier must follow the same behavior. Additionally, the bound is typically loose and does not rule out the

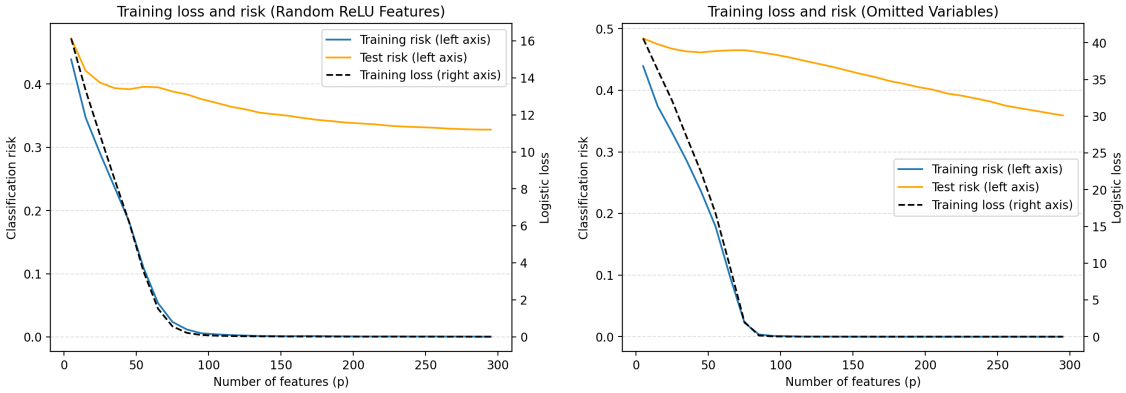

Figure 3: For both random ReLU feature (left) and omitted variables model, weighted logistic loss and classification risk during training decrease monotonically and drop to 0 as number of features increases, and test risk exhibits double descent, with the second descent coinciding with training data interpolation.

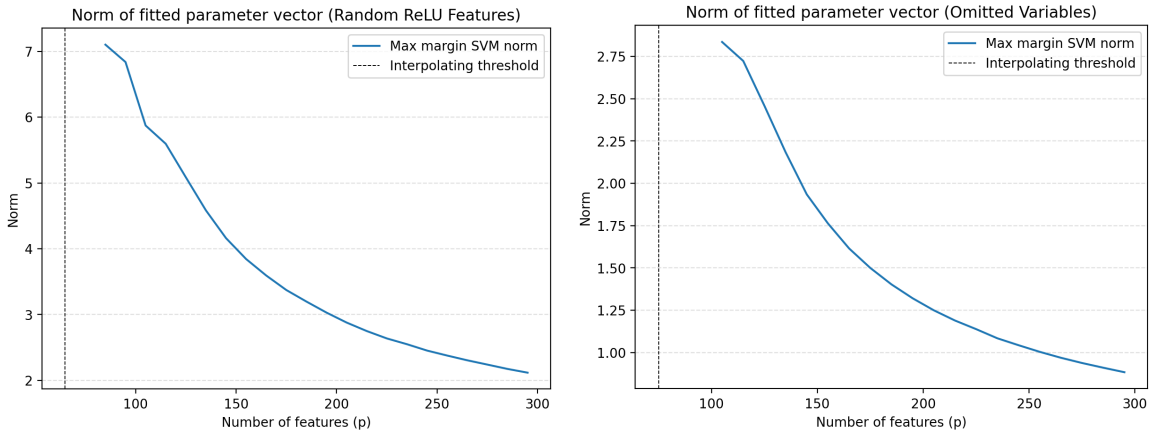

Figure 4: L2 norm of parameter vector from fitting max margin SVM decreases monotonically in the overparameterized regime, indicating larger margin and tighter generalization risk upper bound.

existence of DGP or algorithms under which double descent does not arise. Therefore, this analysis should be interpreted as providing a perspective that is consistent with simulation evidence, rather than a concrete explanation of double descent for general learning methods.

3.2 Direct weighted bounds for linear classifiers

Previously, we derive weighted VC-type generalization bounds for a general class of classifiers. Building on the observation that in the interpolating regime, linear classifier trained with gradient descent converges in direction to the SVM solution, we now present weighted generalization bounds tailored directly to real-valued (mapping to $\mathbb{R}$ instead of $\{+1, -1\}$) linear hypothesis class without replying on first bounding by VC dimensions.

In order to bound real-valued linear functions, we use weighted empirical margin loss instead of 0-1 loss, following Mohri et al. [2018]. The margin loss function is defined as:

$$\Phi_\rho(z) = \begin{cases} 1 \quad if \quad z \leq 0 \\ 1 - \frac{z}{\rho} \quad if \quad 0 \leq z \leq \rho \\ 0 \quad if \quad z > \rho \end{cases} \tag{14}$$

Suppose $z := t_i f(\mathbf{x}_i)$, Φ_ρ is 1 if i is misclassified. But if i is correctly classified with a small margin and its confidence level $t_i f(\mathbf{x}_i)$ falls below the threshold ρ, it still incurs a loss of $1 - \frac{t_i f(\mathbf{x}_i)}{\rho}$. Only predictions with margin exceeding ρ incur zero loss. Thus, ρ controls the confidence required for correct classification. Note that ρ is a hyperparameter set a priori in the margin loss. Then, the weighted empirical margin loss of a function f can be defined as:

$$\hat{R}^\omega_\rho(f) = \frac{1}{n} \sum_{i=1}^{n} \omega_i \Phi_\rho(t_i f(\mathbf{x}_i))$$

The following proposition presents a high probability generalization bound on real-valued linear hypothesis class, assuming bounded weights:

Proposition 3.4 (Weight generalization bound for linear hypothesis) *Suppose x has $||x|| \leq r$ i.e. bounded in a ball of radius r, and weight ω bounded as $0 \leq \omega \leq W$. Let $\mathcal{H}$ be a linear hypothesis class with bounded norm of parameter vector $\{x \mapsto x^T\theta : ||\theta|| \leq \Lambda\}$. Let $\rho > 0$ be a fixed hyperparameter for the margin loss. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size n, the following holds for all $f \in \mathcal{H}$:*

$$R^\omega(f) \leq \hat{R}^\omega_\rho(f) + W\left(2\frac{r\Lambda}{\rho}\sqrt{\frac{1}{n}} + 3\sqrt{\frac{\log 2/\delta}{2n}}\right) \tag{15}$$

$$= O(W\frac{\Lambda}{\rho}\frac{1}{\sqrt{n}}) \tag{16}$$

Proof: Please refer to appendix B.4. □

Next, again relaxing the bounded weights ω assumption and instead assuming sub-Gaussian weights, we have the following corollary:

Corollary 3.2 (Linear hypothesis bound with sub-Gaussian weights) *Suppose the same set-up as Proposition 3.4 except that ω is sub-Gaussian with constant K. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size n, the following holds for all $f \in \mathcal{H}$:*

$$R^{\omega}(f) \leq \hat{R}^{\omega}_{\rho}(f) + K(\sqrt{2\log\frac{2n}{\delta}})\left(2\frac{r\Lambda}{\rho}\sqrt{\frac{1}{n}} + 3\sqrt{\frac{\log 4/\delta}{2n}}\right) \tag{17}$$

$$= O(K\frac{\Lambda}{\rho}\sqrt{\frac{\log n}{n}}) \tag{18}$$

This bound on the class of linear hypothesis directly characterizes the effect of ρ and parameter vector norm Λ on weighted generalization risk. Although one might expect the bounds in this subsection to be tighter than those in the last subsection since the latter apply to any learn method while the former apply to only linear hypothesis, they are not directly comparable due to the presence of the additional hyperparameter ρ and the difference in empirical risk definition.

Comparing the second term of equation 13 to the second term of equation 17, it boils down to comparing $\sqrt{1 + \log(n/||\theta||^2)}$ to $1/\rho$. When sample size n increases, or $||\theta||^2$ decreases with model complexity, $\sqrt{1 + \log(n/||\theta||^2)}$ would become larger than $1/\rho$ and the direct bound on linear hypothesis would yield a tighter uniform upper bound on weighted risk than the bound on general hypothesis class. Setting a larger ρ would also reduce the second term of the bound in 17, but at the cost of larger weighted empirical margin risk since the classifier would need to attain larger confidence for each observation to achieve 0 empirical risk for that observation.

On the other hand, the bound on the norm of the linear hypothesis class Λ has unambiguous effect on the second term of the generalization bound. Fixing ρ, in the overparameterized regime, as feature dimension increases and max-margin SVM searches for interpolating solution in an increasingly smaller linear hypothesis class, Λ decreases and the second term of the generalization bound falls. This leads to a smaller generalization risk upper bound and could be a potential mechanism for the second descent in the overparameterized regime. Gradient descent implicitly regularizes the solution of the linear classifier to a max-margin support vector classifier, resulting in a lower complexity classifier.

4 Further Simulations

4.1 Comparing to shrinkage estimators

When fitting linear model in high dimensional settings, it is natural to think of shrinkage methods. In this section, we compare the performance of ridge and LASSO to unregularized linear classifier.

Regularized linear classifier minimizes the following loss function:

$$\mathcal{L}_r(\theta) := \frac{1}{\sum_{i=1}^{n}\omega_i}\sum_{i=1}^{n}\omega_i \log(1 + \exp(-t_i\mathbf{x}_i^T\theta)) + \lambda\frac{r(\theta)}{\sum_{i=1}^{n}\omega_i} \tag{19}$$

which is weighted logistic loss plus a penalty term on the size of the coefficient vector. For the ridge classifier, $r(\theta) = \frac{1}{2}||\theta||_2^2$, while for the LASSO classifier, $r(\theta) = ||\theta||_1$, similar to regression problem. These methods impose explicit regularization on the loss function while unregularized linear classifier trained with gradient descent undergoes implicit regularization.

We specify two DGPs, which are the same as the random ReLU feature model and the omitted variable model as in section 2.3. Optimal penalty term λ is chosen for ridge and LASSO separately via cross validation on training data, over a grid of λ values, by maximizing out-of-sample welfare (instead of weighted prediction risk). We set training sample size to be 150, test sample size to be 5000, and average results over 100 simulation runs.

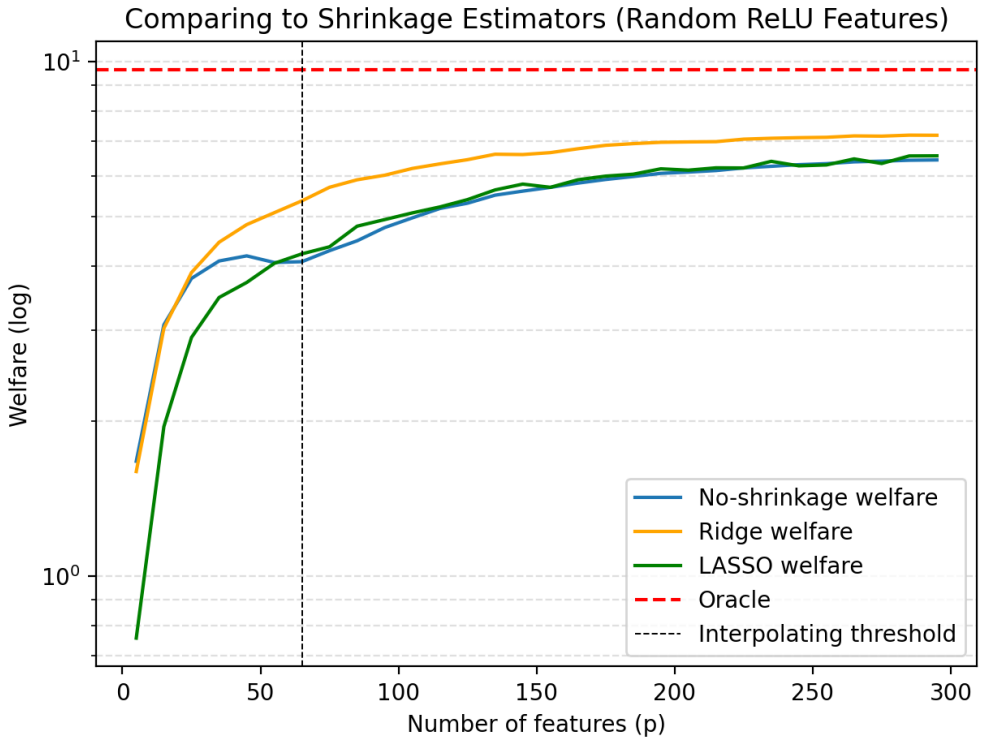

Figure 5: Welfare against the number of features in random ReLU feature expansion. The unregularized linear classifier exhibits double ascnet while ridge and LASSO welfare increases monotonically. Ridge performance dominates across feature dimensions.

Figures 5 and 6 plot test welfare of the three methods against the number of features for random ReLU feature and omitted variables models respectively. Overparameterization improves test welfare for all three methods across the two models. It is notable that for the unregularized classifier, welfare performance dips near the interpolation threshold, but regularized classifiers have welfare growing monotonically in feature dimension, i.e. regularized classifiers do not exhibit double ascent. This is consistent with theoretical findings in Nakkiran et al. [2021b] and Hastie et al. [2022] that, in regression problems, optimally tuned L2-regularization achieves monotonic test performance as sample size or feature dimension changes. Our simulations suggest that under our set-ups, such ridge regression behavior can potentially extend to regularized classification problems, and L1-regularization.

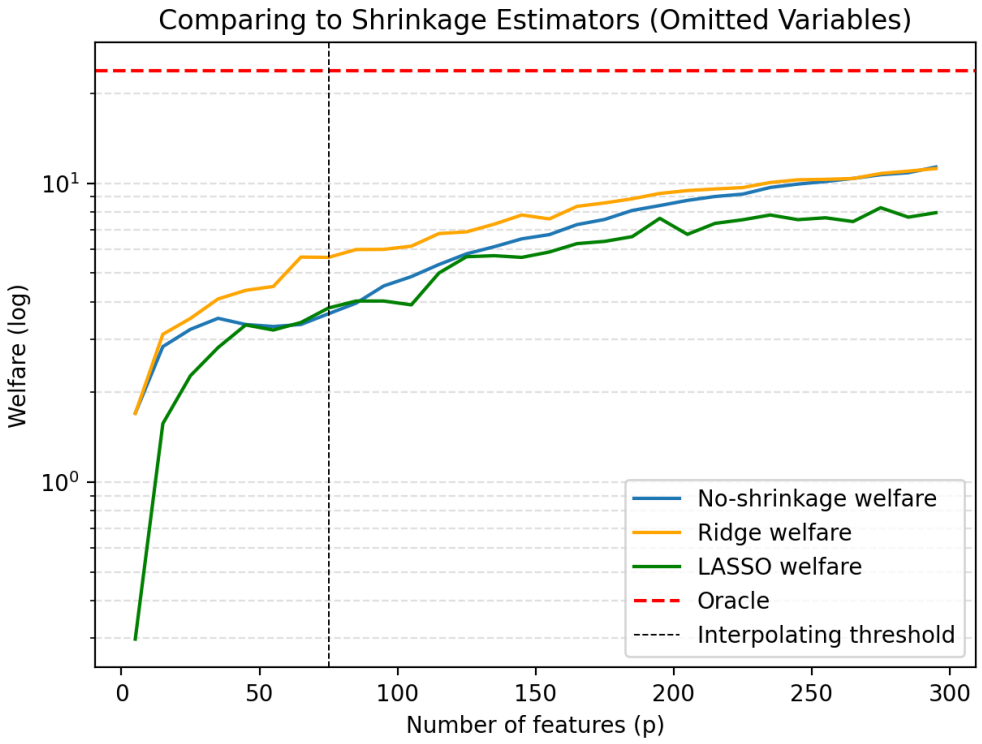

Figure 6: Welfare against the number of features in the omitted variables model. The unregularized linear classifier exhibits double ascent while ridge and LASSO welfare increases monotonically. LASSO underperforms slightly due to imposing sparsity.

In Figure 5, the ridge classifier has better test welfare performance than LASSO and the unregularized classifier across all feature dimensions. This is unsurprising since signal strength is approximately similar for each feature and ridge shrinks the coefficients of all features uniformly. In low dimensional setting, ridge and unregularized classifiers have similar test welfare performance, and outperform LASSO, but ridge dominates the latter as dimensionality increases.

In high dimensional setting, unregularized classifier performs on par with LASSO. This is because LASSO imposes sparsity in the coefficients but there is no sparsity in the random ReLU feature model, hence the optimal LASSO penalty term is chosen to be very small by cross validation. Consider equation 19 with L1-penalty, it has a finite minimizer as long as $\lambda \neq 0$. However, when λ is small, the logistic loss term dominates and the loss minimizing θ can be large. Such loss behavior is similar to the unregularized classifier.

For the omitted variables model, Figure 6 shows qualitatively similar patterns. One noticeable difference is that, under this set-up, in high dimensional setting, the welfare performance of unregularized classifier improves over LASSO and approaches that of the ridge classifier. The difference could be due to that in the ReLU model, each random ReLU feature is formed using all the $\mathbf{x}$'s, but in the omitted variables model, new $\mathbf{x}$'s are included as dimensionality increases which contain more information about the true DGP, hence diminishing the relative performance of shrinkage classifiers.

Taken together, these results suggest that under our set-up, ridge and unregularized classifiers outperform LASSO, and the relative advantage of ridge versus unregularized classifier depends on the true DGP assumptions. In Appendix A, we provide additional simulation results for cases where true DGP changes with dimensionality, which have been commonly studied in past literature.

4.2 Welfare under model misspecification

In this section, we study the behavior of high dimensional sieve estimators under the setting where the true DGP is nonlinear and low-dimensional in $\mathbf{x}$. The goal is to understand double descent behavior when the true DGP is nonlinear and we try to approximate it using a linear model of sieve features. We use the following nonlinear DGP for CATE $\tau(x_1, x_2, x_3)$ with 3 features:

$$\tau(x_1, x_2, x_3) = x_1 + x_1^2 + x_2 + x_2^2 + x_3 + x_3^2 \tag{20}$$

We consider four classes of sieves: polynomial bases, B-spline bases, random ReLU sieves and random Fourier sieves. Polynomial sieve consists of the following terms, up to a total given degree K_1:

$$\phi(\mathbf{x}) = \{x_1^{a_1} x_2^{a_2} x_3^{a_3} : a_1 + a_2 + a_3 \leq K_1\} \tag{21}$$

For B-spline sieve, we construct B-spline basis using a set of K_2 knots placed over the support of the data and a fixed spline degree. The resulting basis functions are piecewise polynomials that are smoothly joined at the knots. We then form the feature map by concatenating the spline expansions across the three coordinates. This creates a flexible and

numerically stable representation that can capture nonlinearities. We set the spline degree to be 3.

Random ReLU features are constructed by:

$$\phi_k(\mathbf{x}) = \max\{a_k^T \mathbf{x} + b_k\} \tag{22}$$

where each element of a_k and b_k is sampled independently from standard normal distribution.

Random Fourier features are constructed by:

$$\phi_k(\mathbf{x}) = \cos(w^T \mathbf{x} + b_k) \tag{23}$$

where each element of w is sampled independently from standard normal distribution and b_k is sampled from $U[0, 2\pi]$.

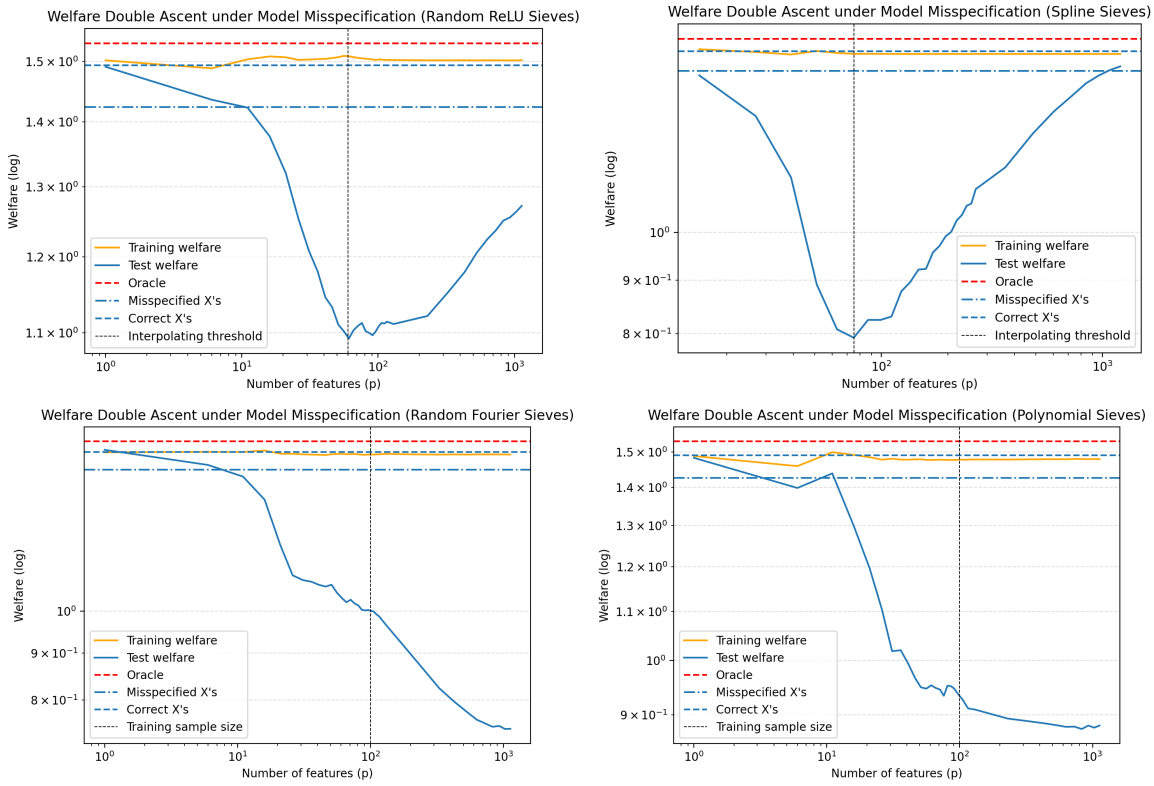

Figure 7: Welfare under model misspecification for four sieve estimators: random ReLU, B-spline, random Fourier and polynomial. The horizontal lines report oracle welfare, welfare from misspecified linear model using the true features, and welfare from the correctly specified model using the true features. Random ReLU and B-spline show welfare improvement in the overparameterized regime, while random Fourier and polynomial do not.

While random ReLU and random Fourier features can be added one by one into the model, it only makes sense for polynomial sieve to grow in the total degree K_1, and B-spline sieve to grow in the total number of knots K_2. We set the training sample size to 100, test sample size to 5000, and the maximum feature dimension to be around 1200 (due

to difference in feature definition and hence feature dimension grows differently across the methods).

Figure 7 shows the test welfare performances of the four sieves against the number of features and reveals sharp contrast between the sieves. Welfare curve labeled "Misspecified X's" plots the welfare obtained from using just x_1, x_2, x_3 in a linear classifier. The curve labeled "Correct X's" plots the welfare obtained from using $x_1, x_1^2, x_2, x_2^2, x_3, x_3^3$ in a linear classifier, which is the true DGP. Thus, we see that "Correct X's" attains higher welfare level than "Misspecified X's".

It is common across all four sieves that in the low dimensional setting, test welfare decreases monotonically in the number of sieve features. Beyond the interplation threshold, random ReLU and B-spline sieves exhibit increasing test welfare. In contrast, test welfare continues decreasing monotonically for random Fourier and polynomial sieves. We find that it is because polynomial features become numerically unstable as degree increases, e.g. polynomial feature raised to power 20 is highly correlated with that raised to power 30. Adding higher degree polynomials does not increase the rank of the design matrix. The same might potentially be the case for random Fourier features, which generate global oscillations that can be correlated when the number of features is large. As such, increasing dimensionality amounts to adding correlated columns and hence does not improve generalization. Such numerical issues do not apply to random ReLU sieves and B-spline sieves.

To understand the difference between the four sieves, we study training loss, training and test risk against the number of features, as shown in Figure 8. For random ReLu and B-spline sieves, training risk quickly falls to 0, indicating that the classifier interpolates training data. For random Fourier and polynomial sieves, interpolation does not happen despite the number of features is much greater than the training sample size. The training risk for random Fourier features appears to be zero but is in fact slightly more than 0. Since there is no interpolation despite the increase in the number of features, our previous analysis about the implicit regularization of a minimum norm interpolator does not apply. In the case of linear classifier, the ability to interpolate training data as the feature dimension increases appears to be necessary for the improvement in test performance in the overparameterized regime.

4.3 Comparing weighted classification to CATE plug-in rule

Two approaches to policy learning are commonly considered. The first estimates CATE and assigns treatment according to the plug-in rule where $\hat{\tau}(\mathbf{x}) > 0$. The second formulates policy learning as a weighted classification problem and directly learns the CATE boundary at zero that maximizes welfare. In this section, we compare the two approaches for the random ReLU model, the omitted variable model, and under model misspecification.

For the plug-in rule, we use T-learner to estimate CATE. First, we fit separate regression models for the treated and control groups by regressing observed outcome y_i on $\mathbf{x}_i$ within each group, and construct CATE as the difference between the two fitted outcome models. Then, individuals with CATE estimate greater than zero are assigned treatment.

We consider settings of both correct specification and misspecification. For correct specification, we use linear DGP with random ReLU feature model and omitted variables model as in Section 2.3. For misspecification, we use the nonlinear DGP as in equation 20,

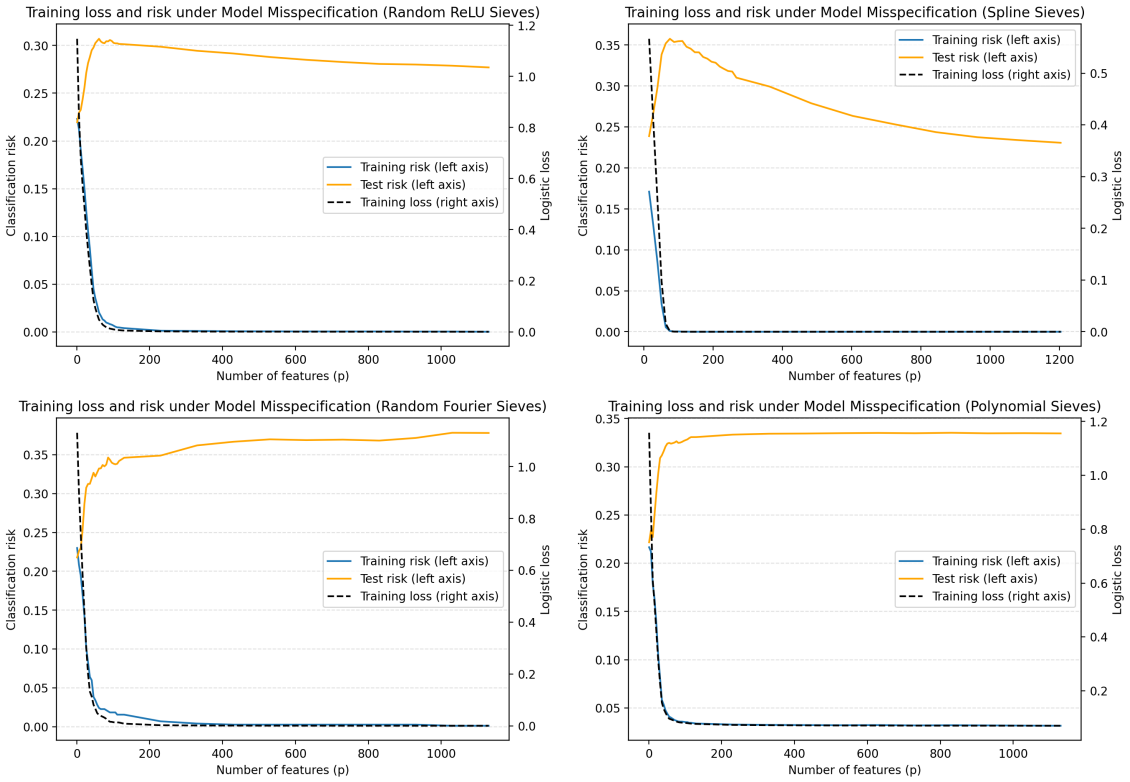

Figure 8: Training loss, training risk and test risk plotted against the number of features. Random ReLU and B-spline show training classification risk reducing to 0 in the number of features, indicating interpolation, while that of random Fourier and polynomial do not go to 0 even as feature dimension increases to around 1200.

with random ReLU and B-spline sieve expansions. We set training sample size to be 150, test sampling size to be 5000, and maximum feature dimension to be 300 in correct specification, and maximum sieve feature dimension to 1200 in misspecification.

Figure 9 plots the results for correct specification, where the left panel shows random ReLU feature model and the right panel shows omitted variables model. In the random ReLU setting, T-learner attains higher test welfare than weighted classification in both heavily underparameterized and overparameterized regimes. The drop in generalization performance of linear regression around the interpolation threshold is more drastic than the drop of weighted classification, hence the latter outperforms for feature dimensions around the interpolation threshold. Note that the interpolation threshold differs for T-learner and weighted classification since the former fits outcome values while the latter learns binary labels. In the overparameterized regime, T-learner achieves welfare much closer to the oracle than weighted classification.

In the omitted variable setting, T-learner again experiences drastic decline in generalization performance near the interpolation threshold, but it outperforms weighted classification in feature dimensions ranging from 100 to 240. Beyond 240, weighted classifier overtakes T-learner in generalization performance. Based on these simulation evidence, there is no clear dominance of one policy learning method over the other.

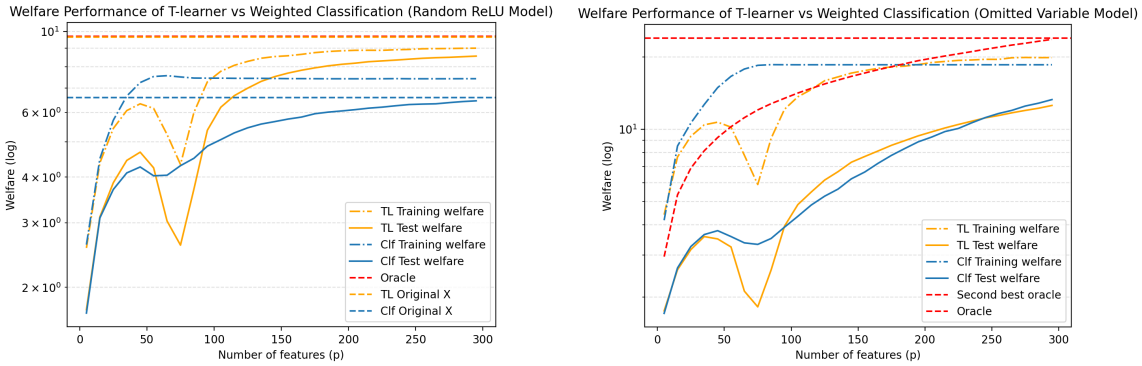

Figure 9: In the left panel, orange is T-learner while blue is weighted classifier. "Original X" welfare refers to using the actual DGP features instead of random ReLU features. Under this setting, T-learner appears to dominate weighted classifier for most parts of feature dimensions except near interpolation threshold where regression generalization performance drops drastically. In the right panel, "second best oracle" refers to the optimal population welfare given access to only p features. It is not obvious whether one method dominates the other in this set up.

Next, under model misspecification, given that the DGP is nonlinear in $\mathbf{x}$, we consider random ReLU and B-spline sieves which have been shown to exhibit double ascent in the previous subsection. Figure 10 ilustrates the welfare performance for the two set-ups. In both cases, T-learner test welfare undergoes a second ascent in the overparameterized regime, but descends again as the sieve feature dimension increases further. The second decrease in performance is much greater when using B-splines than using random ReLU features. On the other hand, weighted classifier test welfare increases monotonically in the overparameterized regime. As such, under our misspecification set-up, weighted classifier demonstrates clear

outperformance over T-learner over nearly the entire range of feature dimensions.

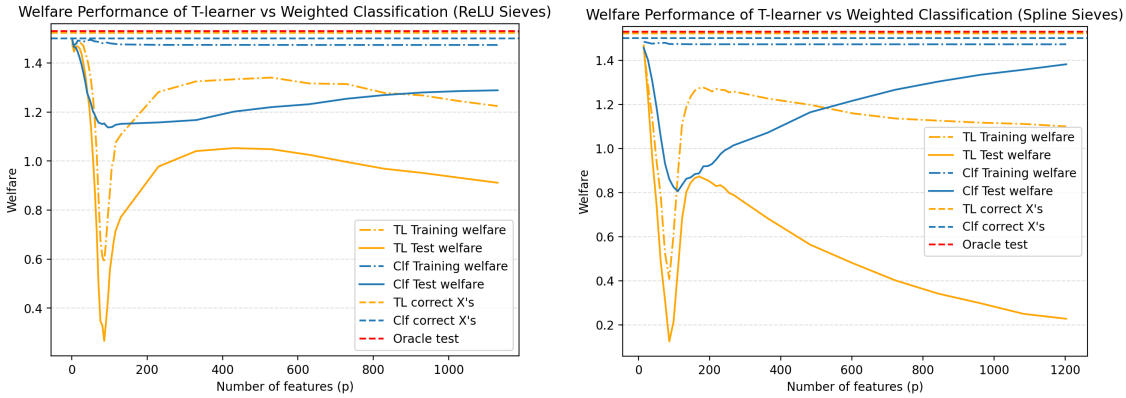

Figure 10: Orange curves illustrate T-learner welfare while blue curves illustrate weighted classifier welfare. For both random ReLU and B-spline sieves, weighted classifier test welfare increases monotonically with feature dimension in the overparameterized regime, while T-learner welfare increases initially when overparameterized, but decreases again as the number of features becomes large.

5 Discussion

This paper documents double ascent in welfare as feature dimension increases in a linear classifier trained with gradient descent using weighted logistic loss. When DGP is linear, under the random ReLU and the omitted variables settings, we consistently observe that highly overparameterized classifier outperforms their low-dimensional counterparts. When DGP is nonlinear and model is misspecified, we provide evidence that sieve estimator experiences improvement in the overparameterized regime only when the classifier interpolates training data and attains zero training risk. The implicit bias induced by gradient descent appears to play a key role in high dimensional setting.

Theoretically, we develop generalization bounds for weighted population risk and hence population welfare by extending VC and margin based analyses to the policy learning setting. These results provide a theoretical perspective for understanding welfare gain in high dimensions. Among the interpolating classifiers, those selected by implicit bias can control complexity, leading to generalization improvement despite higher feature dimension. While such uniform upper bounds are not tight characterization of estimator-specific generalization risk, they offer a perspective for future research.

From a practical perspective, in the modern data-rich environment and increasingly computerized world, decision makers often have access to a large number of features such as high frequency behavioral data. The abundance of information makes it feasible to implement high-dimensional decision rules. At the same time, there is growing acceptance of algorithmic and potentially blackbox decision systems in applications such as personal finance and online platforms which supports the use of complex decision rules. In this context, it might be overly conservative to restrict attention to parsimonious models. Weighted classification provides

a scalable method to include large number of features with the aim of improving welfare objectives.

While we attempt to understand welfare double ascent in policy learning using VC-theoretic framework, our numerous simulations also highlight several important open questions. First, under model misspecification, it remains unclear how to accurately characterize the DGP and class of sieve estimators for which overparameterization improves welfare performance.

Second, we observe that the plug-in approach based on T-learner can outperform weighted classification which learns CATE boundary at zero under correct model specification. Understanding the reason behind the superior performance of T-learner under certain high dimensional settings remains an open question.

Third, the behavior of T-learner based on overparameterized sieve estimator under model misspecification is not well understood. In our simulation, T-learner generalization performance is not monotonic, such as the second welfare decline in high dimensions. Providing a reason for when and why this happens in T-learner but not weighted classifier, e.g. under what type of misspecification and sieve feature expansion, can be a direction for future work.

Four, our analysis assumes either RCT data or that the propensity score is known. This raises the question of how weighted classification performs if the propensity score must be estimated from data in observational settings. Questions include the effect of propensity score estimation error on population welfare and the double ascent pattern.

References

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89: 133–161, 2021.

Peter L. Bartlett, Philip M. Long, Gabor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *PNAS*, 117:30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*, 116:15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571v2*, 2020.

Vladimir Cherkassky and Eng Hock Lee. To understand double descent, we need to understand vc theory. *Neural Networks*, 169:242–256, 2024.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822v2*, 2020.

Yufei Gu, Xiaoqing Zheng, and Tomaso Aste. Unraveling the enigma of double descent: An in-depth analysis through the lens of learned feature space. *arXiv preprint arXiv:2310.13572v3*, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50:949–986, 2022.

Ganesh Ramachandra Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. *ISIT*, 2020.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86:591–616, 2018.

Toru Kitagawa, Shosei Sakaguchi, and Aleksey Tetenov. Constrained classification and policy learning. *arXiv preprint arXiv:2106.12886v2*, 2023.

Eng Hock Lee and Vladimir Cherkassky. Understanding double descent using vc-theoretical framework. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 35, 2024.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *Journal of Machine Learning Research*, 24:1–27, 2023.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72:1221–1246, 2004.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2018.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544v3*, 2023.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics*, 2021a.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897v2*, 2021b.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Neural Information Processing Systems*, 2017.

Daniel Soudry, Elad Hoffer, and Mor Shpigel Nacson. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19:1–57, 2018.

Jann Spiess, Guido Imbens, and Amar Venugopal. Double and single descent in causal inference with an application to high-dimensional synthetic control. *arXiv preprint arXiv:2305.00700v3*, 2023.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2000.

Bianca Zadrozny. *Policy mining: Learning decision policies from fixed sets of data*. PhD thesis, UNIVERSITY OF CALIFORNIA, SAN DIEGO, 2003.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc*, 107:1106–1118, 2012.

Appendix

A Welfare in Isotropic and Factor Models

One common approach in the literature to study the effect of overparameterization is to allow the true DGP to vary with feature dimension while holding the sample size fixed, as in for example Hastie et al. [2022]. In Section 2, we fix a single true DGP and vary feature dimensionality using either random ReLU or omitted variables model. In this section, in contrast, we adopt the alternative approach of allowing DGP to vary with dimension and examine how generalization performance changes.

We suppose CATE is linear in features, and consider two settings of feature distribution. In the first setting, $\mathbf{x}$ is assumed to be isotropic, i.e. independent and same variance, drawn from standard normal distribution. In the second setting, $\mathbf{x}$ is assumed to have a factor model structure as follows:

$$\mathbf{x}_i = F_i B + \nu_i$$

where F_i is a vector of latent factors generated from standard normal distribution and B is a matrix of factor loadings. We set the number of factors to be 25 and the dimension of $\mathbf{x}$ to be 300 for both isotropic and factor $\mathbf{x}$.

For each feature dimension $p \in [1, 300]$, we let CATE be a linear function of $\mathbf{x}_p$ i.e. the true CATE has p features. Crucially, we scale the parameter vector in the true CATE for each p such that signal strength and hence signal to noise ratio remains constant even as feature dimension increases to isolate the effect of overparameterization. We set training sample size to be 150, test sample size to be 5000, and average results over 280 simulation runs.

Figure 11 plots the welfare performance for the isotropic feature model. As the true DGP is changing, it is expected that, fixing training sample size, the model with the highest test welfare is the one with the smallest number of features in the DGP. As feature dimension increases, true DGP becomes more complex but the sample size doesn't increase, resulting in a monotononic decrease in test welfare in the number of features. What is surprising is that beyond the interpolating threshold, although the true DGP is high dimensional versus the training sample size, test welfare improves monotonically in the high dimensional setting.

The same pattern is observed for the factor $\mathbf{x}$ model, except that the magnitude of increase of test welfare in the overparameterized regime is much smaller than using isotropic $\mathbf{x}$, potentially because it is more difficult for the model to learn the true DGP from a very limited training sample when features are correlated.

Under such varying-DGP set-up, we compare the performance of regularized to unregularized linear classifer, with the optimal penalty coefficient chosen by cross-validation. The results are shown in Figures 13 and 14. While ridge dominates LASSO across feature dimension for both istropic and factor features, both ridge and LASSO do not exhibit double ascent. As true DGP becomes more complex, test welfare decreases monotonically. As such, for high dimensional DGPs, unregularized classifier dominates both ridge and LASSO under correct linear model specification.

Next, Figure 15 plots the training loss and training and test risk against feature dimension for both isotropic and factor features. Contrary to the plots in Section 2 which shows

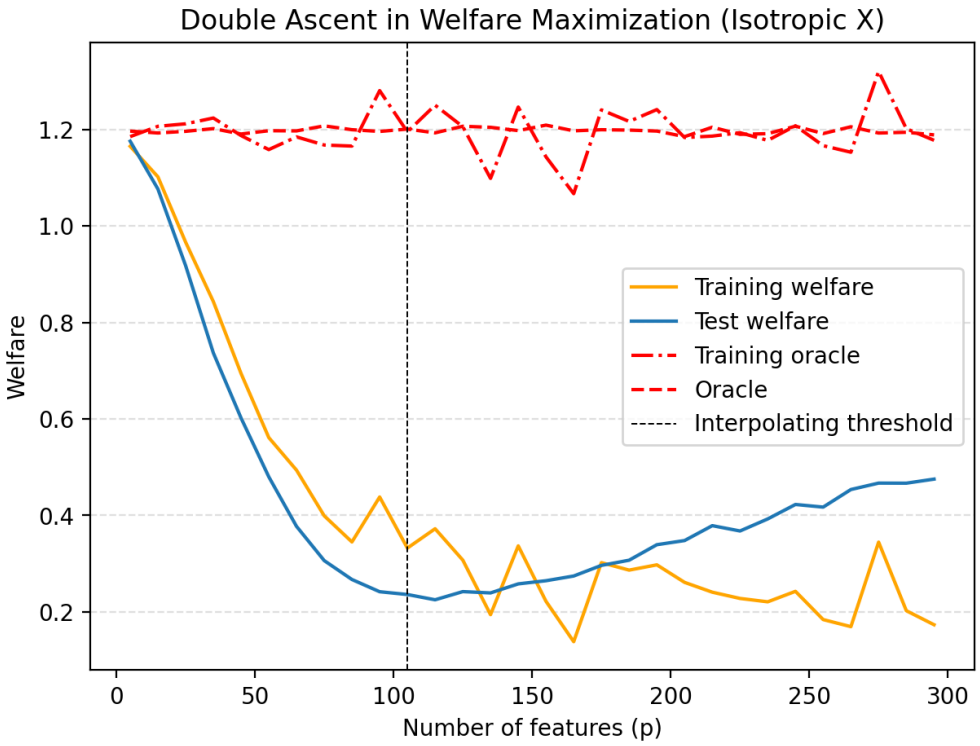

Figure 11: Training oracle is the optimal in-sample welfare that can be obtained given the training sample. Fixing training sample size, test welfare falls initially as DGP becomes more complex. But in the interpolating regime, test welfare rises as the number of features increases. Oracle welfare is constant across feature dimension as we maintain the same signal to noise ratio to isolate the effect of overparameterization.

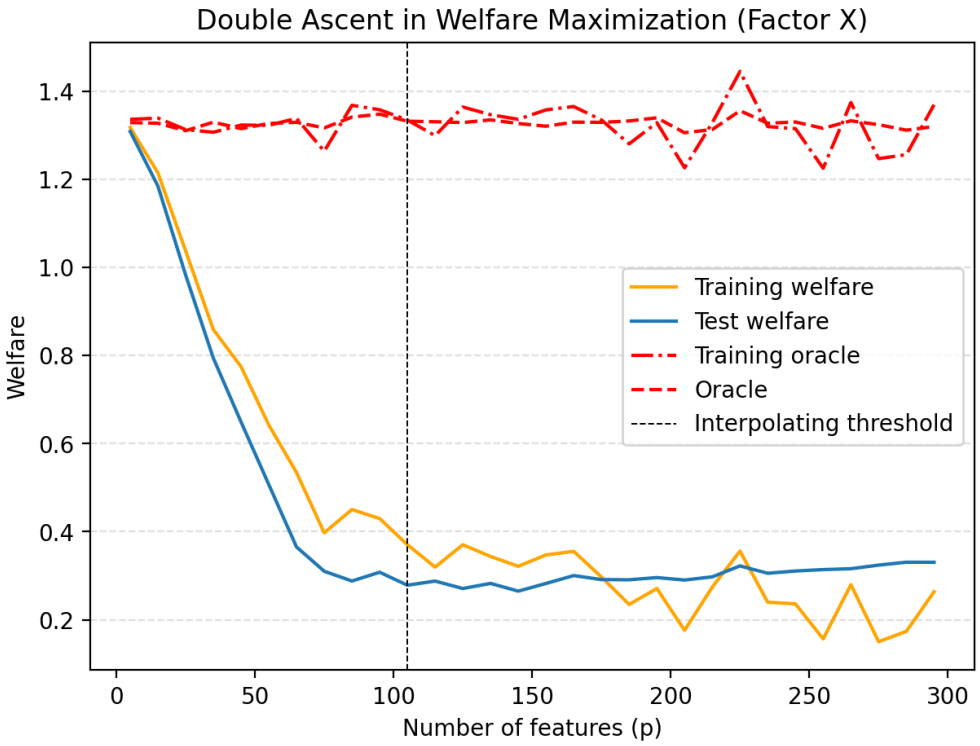

Figure 12: Training oracle is the optimal in-sample welfare that can be obtained given the training sample. When features are generated from a set of low dimensional latent factors and hence correlated, welfare improvement in the overparameterized regime becomes more gradual.

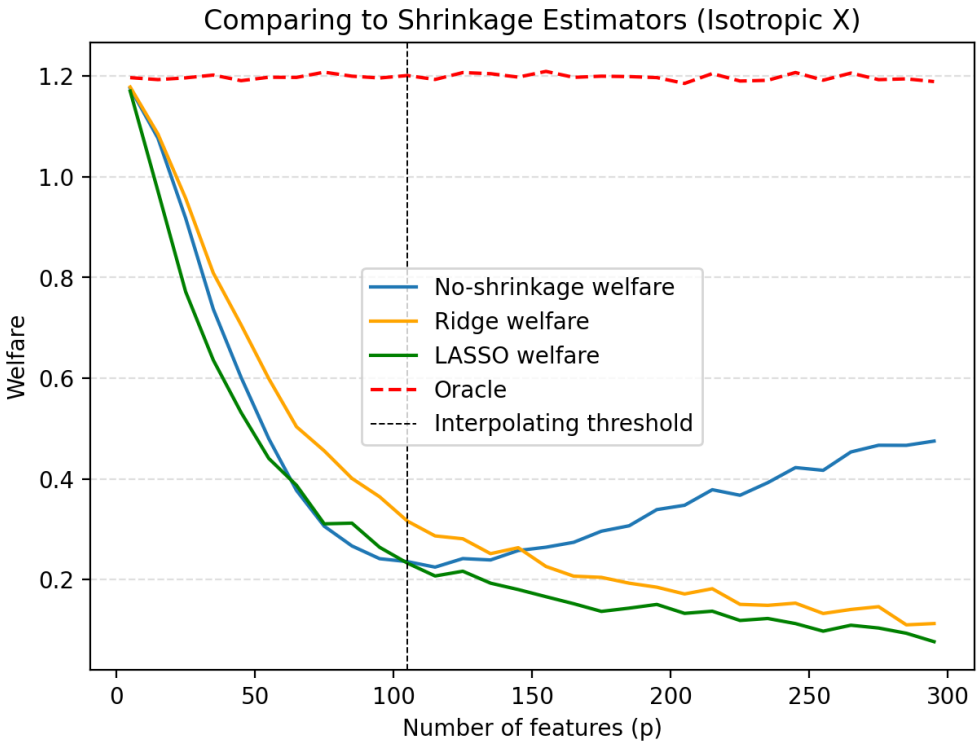

Figure 13: When DGP varies with feature dimension, shrinkage methods no longer display double ascent in welfare, consistent with existing literature. As such, unregularized linear classifier dominates shrinkage methods under correct linear model specification in high dimensional settings.

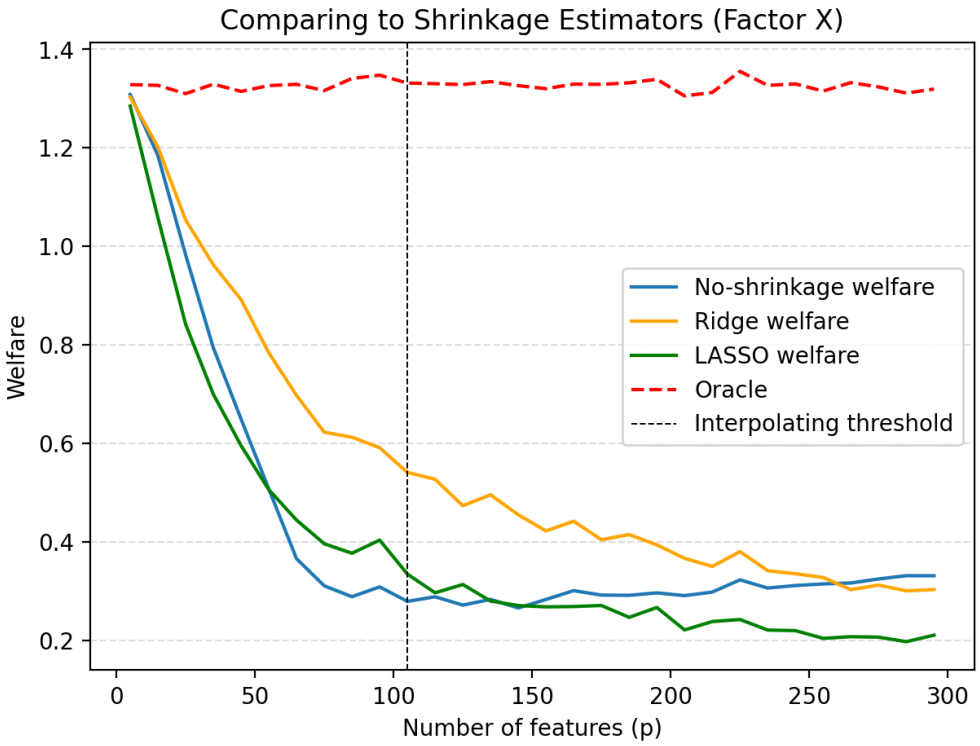

Figure 14: When DGP varies with feature dimension, shrinkage methods no longer display double ascent in welfare, consistent with existing literature. The ascent of test welfare when using factor-structure features is more conspicuous in comparison tho shrinkage methods.

that training risk and loss decrease monotonically under fixed DGP, Figure A shows that training loss and risk rise as DGP becomes more complex and attain a maximum prior to the interpolating threshold, and only decreases to 0 when DGP has enough feature dimension for the fixed-size training sample to be interpolated.

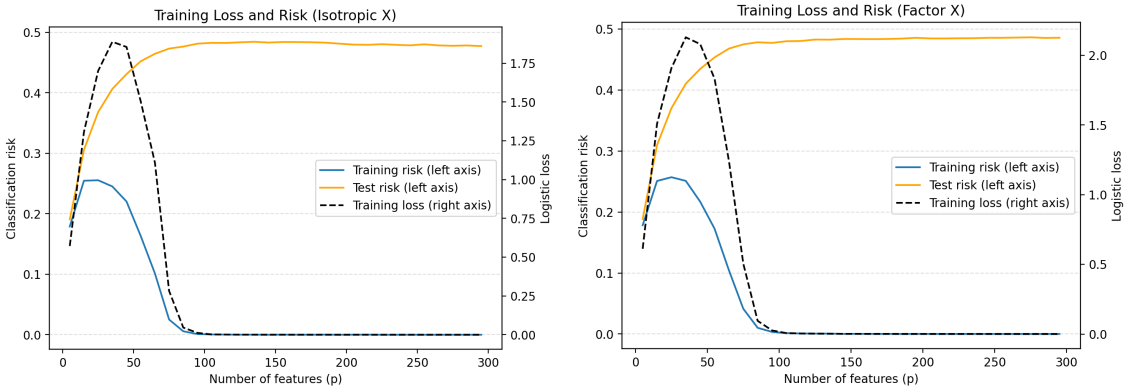

Figure 15: Training loss and risk display very different pattern from the setting where dimensionality varies under a fixed DGP. Instead of decreasing monotonically, training loss and risk peak before the interpolating threshold, and decrease to 0 again, indicating interpolation of training data.

Lastly, Figure 16 shows the norm of the fitted parameter vector in the overparameterized regime. As the feature dimension of DGP increases, fixing sample size, linear classifier trained with gradient descent actually settles in less complex models. It provides evidence that implicit regularization and complexity control in high dimensions can be a potential explanation for the improvement in generalization ability in overparameterized models.

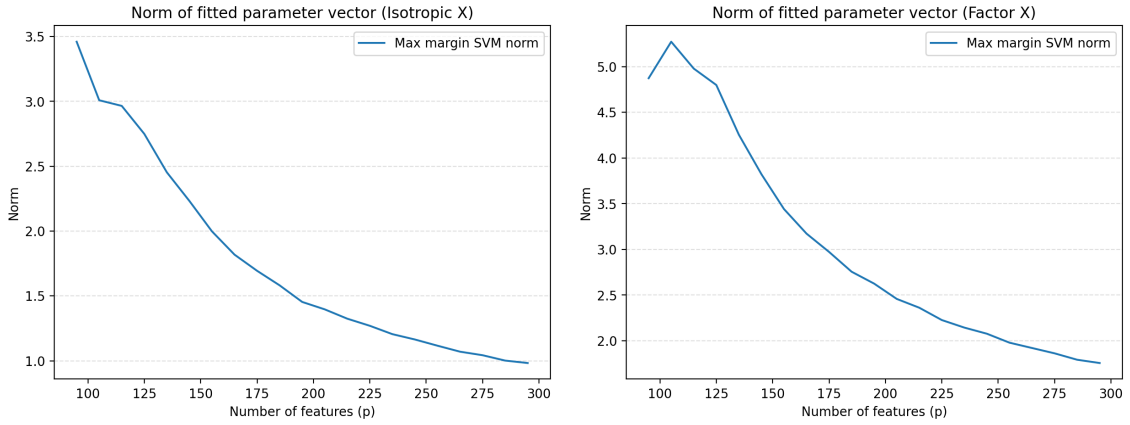

Figure 16: L2 norm of parameter vector from fitting max margin SVM. Under the setting of varying DGP, norm decreases nearly monotonically in the overparameterized regime, implying that larger margin can be obtained for higher feature dimension.

B Proofs for Weighted Generalization Bounds

B.1 Proof for Proposition 3.1

We follow the proof strategy in Mohri et al. [2018] but adapt it to the weighted classification setting. We first bound the empirical process $R^{\omega}(f) - \hat{R}^{\omega}(f)$ uniformly in terms of Rademacher complexity, before bounding Rademacher complexity with VC dimension. We begin by proving the following proposition:

Proposition B.1 *Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{X}$ to $[0,1]$, $\omega(x)$ be a function of $x \in \mathcal{X}$ bounded between $0 \leq \omega \leq W$. Then for any $\delta > 0$, with probability at least $1-\delta$ over an i.i.d sample of S of size n, each of the following holds $\forall g \in \mathcal{G}$:*

$$E[\omega(x)g(x)] \leq \hat{E}_S[\omega(x)g(x)] + W\left(2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}}\right) \tag{24}$$

$$E[\omega(x)g(x)] \leq \hat{E}_S[\omega(x)g(x)] + W\left(2\hat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log 2/\delta}{2m}}\right) \tag{25}$$

where $\hat{E}_S[\omega(x)g(x)] := \frac{1}{n}\sum_{i=1}^{n}\omega(x_i)g(x_i)$ is the empirical average over a sample S, $\hat{\mathfrak{R}}_S(\mathcal{G}) := E_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(x_i)\right]$ is the empirical Rademacher complexity of $\mathcal{G}$ with respect to sample S, and $\mathfrak{R}_n(\mathcal{G}) := E_{S\sim\mathcal{D}^n}[\hat{\mathfrak{R}}_S(\mathcal{G})]$ is the Rademacher complexity of $\mathcal{G}$ over all samples of size n drawn from $\mathcal{D}$.

Proof: Let $\mathcal{V}$ be a class of functions of the form $\{v(x) : v(x) = \omega(x)g(x), g \in \mathcal{G}, 0 \leq \omega \leq W\}$. Note that $v(x)$ is a mapping from $\mathcal{X}$ to $[0, W]$. First, we define the function $\Phi(S) = \sup_{v\in\mathcal{V}}\left(E(v) - \hat{E}_S(v)\right)$ and bound it using McDiarmid's inequality. Consider any two samples S and S' that differ by exactly one point, i.e. x_1 in S and x_1' in S'. The difference between $\Phi(S)$ and $\Phi(S')$ is:

$$\Phi(S) - \Phi(S') \leq \sup_{v\in\mathcal{V}}\left(E(v) - \hat{E}_S(v) - E(v) + \hat{E}_{S'}(v)\right) \tag{26}$$

$$= \sup_{v\in\mathcal{V}}\left(\hat{E}_{S'}(v) - \hat{E}_S(v)\right) \tag{27}$$

$$= \sup_{v\in\mathcal{V}}\frac{v(x_1') - v(x_1)}{n} \tag{28}$$

$$\leq \frac{W}{n} \tag{29}$$

We can obtain $\Phi(S') - \Phi(S) \leq \frac{W}{n}$ in the same way, hence $|\Phi(S) - \Phi(S')| \leq \frac{W}{n}$. Then, applying McDiarmid's inequality, for all $\epsilon > 0$, the following holds:

$$\mathbb{P}[(\Phi(S) - E[\Phi(S)]) \geq \epsilon] \leq \exp(\frac{-2\epsilon^2 n}{W^2})$$

and

$$\mathbb{P}[(\Phi(S) - E_S[\Phi(S)]) \leq \epsilon] \geq 1 - \exp(\frac{-2\epsilon^2 n}{W^2})$$

Equating the RHS to $1-\delta$, we get

$$\epsilon = \sqrt{\frac{W^2}{2n}\log(\frac{1}{\delta})}$$

Hence, for any $\delta > 0$, with probability of at least $1-\delta$, we have:

$$\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{W^2}{2n}\log(\frac{1}{\delta})} \tag{30}$$

Next, we bound $E_S[\Phi(S)]$ by the following:

$$\begin{aligned}
E_S[\Phi(S)] &= E_S[\sup_{v\in\mathcal{V}}(E(v) - \hat{E}_S(v))] && (31)\\
&= E_S[\sup_{v\in\mathcal{V}} E_{S'}[\hat{E}_{S'}(v) - \hat{E}_S(v)]] && (32)\\
&\leq E_{S,S'}[\sup_{v\in\mathcal{V}}(\hat{E}_{S'}(v) - \hat{E}_S(v))] && (33)\\
&= E_{S,S'}[\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}(v(x_i') - v(x_i))] && (34)\\
&= E_{\sigma,S,S'}[\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(v(x_i') - v(x_i))] && (35)\\
&\leq E_{\sigma,S'}[\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(x_i')g(x_i')] + E_{\sigma,S}[\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i w(x_i)g(x_i)] && (36)\\
&\leq E_{\sigma,S'}[\sup_{g\in\mathcal{G}}\frac{W}{n}\sum_{i=1}^{n}(\sigma_i g(x_i'))] + E_{\sigma,S}[\sup_{g\in\mathcal{G}}\frac{W}{n}\sum_{i=1}^{n}(\sigma_i g(x_i))] && (37)\\
&\leq 2WE_{\sigma,S}[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(x_i)] && (38)\\
&= 2W\mathfrak{R}_n(\mathcal{G}) && (39)
\end{aligned}$$

This proves the first inequality in the proposition. To prove the second inequality, we can use McDiarmid's inequality to bound $\hat{\mathfrak{R}}_S(\mathcal{G})$ to obtain:

$$\mathbb{P}[(\hat{\mathfrak{R}}_S(\mathcal{G}) - \mathfrak{R}_n(\mathcal{G})) \leq -\epsilon] \leq \exp(-2n\epsilon^2)$$

Hence with probability $1-\delta/2$, we have:

$$\mathfrak{R}_n(\mathcal{G}) \leq \hat{\mathfrak{R}}_S(\mathcal{G}) + \sqrt{\frac{1}{2n}\log(\frac{2}{\delta})}$$

Then, we bound $\Phi(S) - E_S[\Phi(S)]$ in the same way as before but using $\delta/2$ instead of δ, so we have:

$$\Phi(S) \leq 2W\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{W^2}{2n}\log(\frac{2}{\delta})} \tag{40}$$

$$\leq 2W\hat{\mathfrak{R}}(\mathcal{G}) + 3W\sqrt{\frac{1}{2n}\log(\frac{2}{\delta})} \tag{41}$$

Because we apply McDiarmid's inequality twice, each time with probability $1-\delta/2$, by union bound, this gives us the second inequality in the proposition with probability at least $1-\delta$. □

After showing Proposition B.1, the remainder of the proof for Proposition 3.1 follows the same strategy as Mohri et al. [2018], which bounds the Rademacher complexity of $\mathcal{G}$ by the VC dimension of $\mathcal{H}$. Let $\mathcal{H}$ be a class of functions mapping from $\mathcal{X}$ to $[+1, -1]$, and $\mathcal{G}$ be a class of loss functions mapping to $[0, 1]$ i.e. $\mathcal{G} := \{\mathcal{X} \times \{+1, -1\} \mapsto \mathbf{1}\{f(x) \neq y\} : f \in \mathcal{H}, x \in \mathcal{X}, y \in \{+1, -1\}\}$. Then, Proposition B.1 states that:

$$R^{\omega}(f) \leq \hat{R}^{\omega}(f) + W\left(2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right)$$

Next, we bound the Rademacher complexity of $\mathcal{G}$ in terms of the VC dimension of $\mathcal{H}$:

$$R^{\omega}(f) \leq \hat{R}^{\omega}(f) + W\left(2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \tag{42}$$

$$= \hat{R}^{\omega}(f) + W\left(\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \quad \text{(Lemma 3.4)} \tag{43}$$

$$\leq \hat{R}^{\omega}(f) + W\left(\sqrt{\frac{2\log\Pi_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \quad \text{(Massart's Lemma \& Corollary 3.8)} \tag{44}$$

$$\leq \hat{R}^{\omega}(f) + W\left(\sqrt{\frac{2d\log(\frac{en}{d})}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \quad \text{(Sauer's Lemma \& Corollary 3.18)} \tag{45}$$

where $\mathfrak{R}_n(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$, $\Pi_{\mathcal{H}}(n)$ is the growth function for $\mathcal{H}$ with m sample points, and d is the VC dimension of $\mathcal{H}$. Note that the lemmas and corollaries that we use to establish the above inequalities can be found in Mohri et al. [2018]. This completes the proof for Proposition 3.1.

B.2 Proof for Proposition 3.2

Proposition 3.2 assumes sub-Gaussian instead of unbounded weights. To prove Proposition 3.2, we adopt the strategy of conditioning on a high probability event where the weights are uniformly bounded on the sample, proving the weighted generalization bound, and finally dropping the conditioning by a union bound. By the definition of sub-Gaussian,

$$\mathbb{P}(w(x) \geq t) \leq \exp(-\frac{t^2}{2K^2})$$

where K is the sub-Gaussian constant. Then, for a sample of size n, define the event that the weights are uniformly bounded by W as:

$$\mathcal{E}_W := \{\max_{1\leq i\leq n} w_i \leq W\}$$

where $w_i = w(x_i)$. Then its complement is:

$$\mathcal{E}_W^C := \{\exists i, w_i > W\}$$

To have probability at least $1-\delta_0$ that the weights are uniformly bounded by W, we first bound the probability of $\mathcal{E}_W^C$ and set it equal to δ_0:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_W^C) &= \mathbb{P}(\{\exists i, w_i > W\}) && (46)\\
&= \mathbb{P}(\cup_{i=1}^n \{w_i > W\}) && (47)\\
&\leq \sum_{i=1}^n \mathbb{P}(w_i > W) && (48)\\
&\leq n\exp(-\frac{W^2}{2K^2}) && (49)
\end{aligned}$$

By setting $W = K\sqrt{2\log\frac{n}{\delta_0}}$, we have that the weighted are uniformly bounded by W with probability at least $1-\delta_0$. Then, conditional on $\mathcal{E}_W$, we derive the weighted generalization bound in terms of W just as we did for the proof of Proposition 3.1, with probability at least $1-\delta_1$:

$$R^\omega(f) \leq \hat{R}^\omega(f) + W\left(\sqrt{\frac{2d\log(\frac{en}{d})}{n}} + \sqrt{\frac{\log\frac{1}{\delta_1}}{2n}}\right)$$

Then, choosing $\delta_0 = \delta_1 = \frac{1}{2}$, and by union bound, we have that for any $\delta > 0$, the following holds with probability at least $1-\delta$:

$$\begin{aligned}
R^\omega(f) &\leq \hat{R}^\omega(f) + K\sqrt{2\log\frac{2n}{\delta}}\left(\sqrt{\frac{2d\log(\frac{en}{d})}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right) && (50)\\
&= \hat{R}^\omega(f) + K\sqrt{\log\frac{2n}{\delta}}\left(2\sqrt{\frac{d\log\frac{en}{d}}{n}} + \sqrt{\frac{1}{n}\log\frac{2}{\delta}}\right) && (51)\\
&= O\left(\sqrt{\log n}\sqrt{\frac{\log\frac{n}{d}}{\frac{n}{d}}}\right) && (52)
\end{aligned}$$

B.3 Proof for Corollary 3.1

Proof: Following Proposition 2.1, empirical welfare can be expressed as:

$$\hat{W}(f) = \frac{1}{n}\sum_{i=1}^n \left(\max\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\}\right)$$

$$- \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i|}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\mathbf{1}\{sign(y_i)\cdot D_i \cdot sign(f(\mathbf{x_i})) < 0\}\right) \tag{53}$$

The remainder of the proof is just algebra, with the inequality following Proposition 3.2:

$$\begin{aligned} W(f) - \hat{W}(f) = E_Q[\max\left\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\right\}] - \frac{1}{n}\sum_{i=1}^{n}\left(\max\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\}\right) \\ - E_Q[\frac{|y_i|}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\mathbf{1}\{sign(y_i)\cdot D_i \cdot sign(f(\mathbf{x_i})) < 0\}] \\ + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i|}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\mathbf{1}\{sign(y_i)\cdot D_i \cdot sign(f(\mathbf{x_i})) < 0\}\right) \end{aligned} \tag{54}$$

$$\begin{aligned} \geq E_Q[\max\left\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\right\}] - \frac{1}{n}\sum_{i=1}^{n}\left(\max\{0, \frac{y_i}{D_i e(\mathbf{x_i}) + (1-D_i)/2}\}\right) \\ - K\sqrt{\log\frac{2n}{\delta}}\left(2\sqrt{\frac{d\log\frac{en}{d}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}\right) \end{aligned} \tag{55}$$

□

B.4 Proof for Proposition 3.4

The proof follows the same strategy as that in Mohri et al. [2018]. Let $\mathcal{H}$ be a class of real-valued functions, let $z_i = (x_i, t_i)$ where $t_i \in \{+1, -1\}$, and define $\mathcal{G}$ as:

$$\mathcal{G} := \{z_i \mapsto t_i f(x_i) : f \in \mathcal{H}\}$$

and define $\tilde{\mathcal{G}}$ as the class of functions compositing $\Phi_\rho(.)$ with g for $g \in \mathcal{G}$:

$$\tilde{\mathcal{G}} := \{z_i \mapsto \Phi_\rho \circ g(z_i) : g \in \mathcal{G}\}$$

Since any function in $\tilde{\mathcal{G}}$ is a mapping to $[0, 1]$, by Proposition B.1, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for each $\tilde{g} \in \tilde{\mathcal{G}}$:

$$E[w(z)\tilde{g}(z)] \leq \hat{E}[w(z)\tilde{g}(z)] + W\left(2\hat{\mathfrak{R}}_S(\tilde{\mathcal{G}}) + 3\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

and writing this in terms of $f \in \mathcal{H}$:

$$E[w(z)\Phi_\rho(t\cdot f(x)] \leq \hat{R}^{\omega}_{\rho}(f) + W\left(2\hat{\mathfrak{R}}_S(\Phi_\rho \circ \mathcal{G}) + 3\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

By Lemma 5.7 (Talagrand's lemma) in Mohri et al. [2018], since Φ_ρ is $\frac{1}{\rho}-$Lipschitz and $\mathcal{G}$ is a hypothesis set of real-valued functions, we have that $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \mathcal{G})$ is bounded by:

$$\hat{\mathfrak{R}}_S(\Phi_\rho \circ \mathcal{G}) \leq \frac{1}{\rho}\hat{\mathfrak{R}}_S(\mathcal{G})$$

and hence we have:

$$E[w(z)\Phi_\rho(t \cdot f(x)] \leq \hat{R}^\omega_\rho(f) + W\left(\frac{2}{\rho}\hat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

Moreover, the Rademacher complexity of $\mathcal{G}$ is equal to that of $\mathcal{H}$:

$$\hat{\mathfrak{R}}_S(\mathcal{G}) := E_\sigma[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i t_i f(x_i)] = E_\sigma[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)] = \hat{\mathfrak{R}}_S(\mathcal{H})$$

The second inequality holds because multiplying σ_i by a binary label y_i does not change its distribution. This, coupled with the fact that $\Phi_\rho(z)$ upper bounds $\mathbf{1}(z \leq 0)$, we have:

$$R^\omega(f) \leq \hat{R}^\omega_\rho(f) + W\left(\frac{2}{\rho}\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

Finally, by Theorem 5.10 of Mohri et al. [2018] which states that for vector x with $||x|| \leq r$ bounded in a ball of radius r and sample size n, and for $\mathcal{H} := \{x \mapsto x^T\theta : ||\theta|| \leq \Lambda\}$, the empirical Rademacher complexity of $\mathcal{H}$ is bounded by:

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{n}}$$

We end up with the bound on weighted population risk in terms of $||\theta||$:

$$R^\omega(f) \leq \hat{R}^\omega_\rho(f) + W\left(\frac{2}{\rho}\sqrt{\frac{r^2\Lambda^2}{n}} + 3\sqrt{\frac{\log 2/\delta}{2n}}\right) \tag{56}$$

$$= \hat{R}^\omega_\rho(f) + W\left(2\frac{r\Lambda}{\rho}\sqrt{\frac{1}{n}} + 3\sqrt{\frac{\log 2/\delta}{2n}}\right) \tag{57}$$

This completes the proof for Proposition 3.4.

B.5 Proof for Corollary 3.2

The proof for this corollary follows the same strategy as that for Proposition B.2. We condition on a high probability event where the weights are uniformly bounded on the sample, proving the generalization bound, and finally dropping the conditioning.

Following the same steps as in the section B.2, we set $W = K\sqrt{2\log\frac{n}{\delta_0}}$ and choose $\delta_0 = \delta_1 = \frac{\delta}{2}$, giving us that for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$R^\omega(f) = \hat{R}^\omega_\rho(f) + K\sqrt{2\log\frac{2n}{\delta_0}}\left(2\frac{r\Lambda}{\rho}\sqrt{\frac{1}{n}} + 3\sqrt{\frac{\log 4/\delta}{2n}}\right) \tag{58}$$

C Proof for Proposition 3.3

Soudry et al. [2018]'s proof for convergence in direction of linear classifier trained with gradient descent to max-margin SVM cannot be applied mechanically in our case because their results apply to empirical loss of the form $\sum_{i=1}^{n} \ell(y_n \mathbf{x}_i^T \theta)$ where each unit has the same functional form of loss, while our weighted empirical loss function is unit-specific i.e. each unit has loss $\omega_i \ell(y_n \mathbf{x}_i^T \theta)$. Our proof strategy is to first verify that each unit-specific loss satisfies the loss function assumptions required by Soudry et al. [2018], followed by showing that their main arguments continue to hold for the weighted loss objective.

First, Soudry et al. [2018] Assumption 2 requires $\forall u, \ell(u) > 0, \ell'(u) < 0, \lim_{u\to\infty} \ell(u) = \lim_{u\to\infty} \ell'(u) = 0$, $\ell'(u)$ is β-Lipschitz and $\lim_{u\to-\infty} \ell'(u) \neq 0$. Assumption 3 requires that $-\ell'(u)$ has a tight exponential tail. Multiplying $\ell(u)$ by a constant ω_i that is bounded away from 0 and infinity does not violate these assumptions i.e. these conditions hold true for each unit-specific loss. Following their convention, we redefine $y_i \mathbf{x}_i^T \theta$ by setting $\mathbf{x}_i \leftarrow \mathbf{x}_i y_i$ i.e. multiplying $\mathbf{x}_i$ by y_i and setting all y_i to have label $+1$, WLOG, e.g with it, linear separability can be defined as there exists a θ_* such that $\mathbf{x}_i^T \theta_* > 0$.

Next, we show that Soudry et al. [2018] Lemma 1 holds true, i.e. let $\theta(t)$ be the GD parameter at iteration t with learning rate $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$, we have that as $t \to \infty$, weighted loss $\mathcal{L}^\omega := \sum_{i=1}^{n} \omega_i \ell(y_n \mathbf{x}_i^T \theta)$ converges to 0, $||\theta||$ converges to infinity, and $\mathbf{x}_i^T \theta(t)$ converges to infinity for all i. Since data is separable, there exists a θ_* such that $\theta_*^T \mathbf{x}_i > 0$ for all i and

$$\theta_*^T \nabla \mathcal{L}^\omega(\theta) = \sum_{i=1}^{n} \omega_i \ell'(\mathbf{x}_i^T \theta) \theta_*^T \mathbf{x}_i < 0$$

since $\ell'(u) < 0$ and ω_i are strictly positive. Therefore, $\nabla \mathcal{L}^\omega(\theta)$ cannot be exactly 0 for finite θ. However, running gradient descent on a smooth loss function with appropriate stepsize is guaranteed to converge to a critical point with zero gradient. This implies that $\theta(t)$ diverges to infinity in order for $\nabla \mathcal{L}^\omega(\theta)$ to go to 0. Also, $\theta(t)^T \mathbf{x}_i$ must diverge to infinity in order for $\ell'(\theta(t)^T \mathbf{x}_i)$ to go to 0. Then, by Assumption 2, $\ell(\theta(t)^T \mathbf{x}_i)$ goes to 0 and $\mathcal{L}^\omega(\theta)$ goes to 0. Hence Lemma 1 holds for weighted loss.

With that, we can then show that Soudry et al. [2018]'s Theorem 9 holds true under weighted loss and hence establishes Proposition 3.3. First, assuming for simplicity that $\ell(u) = e^{-u}$. If $\frac{\theta(t)}{||\theta(t)||}$ converges to some limit θ_∞, then we can write $\theta(t) = g(t)\theta_\infty + \rho(t)$ where $g(t) \to \infty$, $\mathbf{x}_i^T \theta_\infty > 0 \ \ \forall i$, and $\lim_{t\to\infty} \frac{\rho(t)}{g(t)} \to 0$. Note that $\rho(t)$ is a vector with the same dimension as $\theta(t)$. The gradient flow can be written as:

$$\nabla \mathcal{L}^\omega(\theta) = \sum_{i=1}^{n} \omega_i \exp(-\theta(t)^T \mathbf{x}_i) \mathbf{x}_i = \sum_{i=1}^{n} \omega_i \exp(-g(t)\theta_\infty^T \mathbf{x}_i) \exp(-\rho(t)^T \mathbf{x}_i) \mathbf{x}_i$$

As $g(t)$ goes to infinity, $-g(t)\theta_\infty^T \mathbf{x}_i$ becomes more negative and the exponent approaches 0. Observations with larger $\theta_\infty^T \mathbf{x}_i$ have exponents going to 0 more quickly and only observations with smaller $\theta_\infty^T \mathbf{x}_i$ will contribute to the gradient. These are the observations with the smallest margins i.e. the support vectors. Pre-multiplying the loss derivative by a constant weight ω_i does not change this behavior since the change in exponent quickly dominates the scale of ω_i. As such, when t increases, as the exponent term of non-support vector goes to 0

quickly, gradient will depend mostly on the support vectors and the limit θ_∞ will be a linear combination of support vectors:

$$\hat{\theta} := \frac{\theta_\infty}{\min_i \theta_\infty^T \mathbf{x_i}} = \sum_{i=1}^{n} a_i \mathbf{x}_i$$

and, for all i, it is either

$$(a_i \geq 0 \text{ and } \hat{\theta}^T \mathbf{x}_i = 1) \text{ or } (a_i = 0 \text{ and } \hat{\theta}^T \mathbf{x}_i > 1)$$

with the former case corresponding to the support vectors. Therefore, with weighted loss, we also arrive at the same KKT conditions for the unweighted hard margin SVM problem, and we can conclude that the direction of the gradient descent estimate converges to that of the SVM solution.

Finally, it remains to show that the limit $\frac{\theta(t)}{||\theta(t)||}$ exists, extend the form of loss we assumed earlier to the general tight exponential tail loss, and most crucially prove that the residual error of gradient descent is bounded. Soudry et al. [2018] provide details on these in their Appendix A.2. It is straightforward to go through their calculation to see that their conclusions for unweighted loss carry through and $\rho(t)$ remains bounded by $C_1 t^{-v}$ for some $C_1 > 0$ and $v > 1$ when the loss function is pre-multiplied by a constant factor ω_i. Therefore, their arguments for Theorem 9 remain valid for weighted empirical loss, and the conclusion of convergence in direction to the max margin SVM carries through to weighted loss function for all datasets with non-zero measure. We postulate that multiplying weights in the loss function affects the convergence rate but not the limit to max margin SVM, and the effect depends on the relative magnitudes of weights on the support and non-support vectors.