The Equivalence of Robustness and Regularization with Non-Perturbable Predictors

Sheng Yang*

December 9, 2022

1 Introduction

In machine learning, uncertainty in the dataset has a huge impact on its performance. Regularization is the most common heuristic approach to alleviate such influences. Recent works show that there is an exact equivalence between robustness against uncertainty and regularization in several contexts (El Ghaoui and Lebret, 1997; Bertsimas and Copenhaver, 2018) and later in the most general form (Ben-Tal et al., 2009; Bertsimas and Dunn, 2019).

However, previous works prove such equivalence assuming all features are perturbable (i.e. every feature admit a continuous perturbation). This assumption does not hold in general, since dataset may contain non-continuously perturbable features: categorical features such as gender are discrete variables and it is unreasonable to guard this features against continuous noise; non-perturbable continuous features are also prevalent, particularly in medical trials where the drug level administered to individuals are predetermined and does not contain random components.

In this project we show that the equivalence of robustness and regularization with non-perturbable features still holds. Specifically, this finding can be applied to robust classification (Bertsimas et al., 2019) to make robust logisitc regresssion, robust support vector machine, and optimal classification tree applicable in a broader context.

This report is organized in the following structure: we first introduce the basic form of such equivalence, and then provide theorems and proofs for their counterparts with non-perturbable features in the general contexts. Lastly we discuss its application in robust classification, followed by synthetic and real world empirical studies to demonstrate its strength and limitations.

2 Equivalence with Non-perturbable Features: General Form

Recall from (Bertsimas and Dunn, 2019) chapter 2 the following theorem:

Theorem. *If $g : \mathbb{R}^n \mapsto \mathbb{R}$ is a seminorm, which is not identically zero and $h : \mathbb{R}^n \mapsto \mathbb{R}$ is a norm, then for any $\boldsymbol{z} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$,*

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(h,g)}} g(\boldsymbol{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) = g(\boldsymbol{z}) + \lambda h(\boldsymbol{\beta}) \tag{1}$$

*Harvard John Paul School of Engineering and Applied Science, shengyang@g.harvard.edu

where $\mathcal{U}_{(h,g)} = \{\boldsymbol{\Delta} : ||\boldsymbol{\Delta}||_{(h,g)} \leqslant \lambda\}$.

Note that $\boldsymbol{\Delta\beta}$ refers to the robust portion, and $\lambda h(\boldsymbol{\beta})$ is the regularization. All estimated parameters are regularized if every feature is continuously perturbable. A wishful thinking may suggest that only the parameters corresponding to perturbable feature part are regularized if only a subset of features are continuously pertrubable.

This is indeed true by making the following observation: consider a dataset $X \in \mathbb{R}^{n\times p}$ with k non-perturbable features, where $0 \leqslant k \leqslant p$. We could always rearrange columns so that the last k features are non-perturbable. Instead of restricting $\boldsymbol{\Delta}$ towards our purpose, we limit $\boldsymbol{\beta'}$ to a particular subspace of $\mathbb{R}^p$ where only the first $p-k$ coordinates are allowed non-zero. Note that this could be denoted by $\boldsymbol{\beta'} = \mathbf{D}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\mathbf{D}$ is a diagonal matrix with the first $p-k$ diagonal entries being 1s and the other 0s.

Substituting $\boldsymbol{\beta'}$ with $\mathbf{D}\boldsymbol{\beta}$ in Equation (1), we have

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{(h,g)}} g(\boldsymbol{z} + \boldsymbol{\Delta}\mathbf{D}\boldsymbol{\beta}) = g(\boldsymbol{z}) + \lambda h(\mathbf{D}\boldsymbol{\beta}) \tag{2}$$

This fulfills our purpose since $\boldsymbol{\Delta}\mathbf{D}$ only keeps the perturbable permutations and silences the part for the non-perturbable features, and $\mathbf{D}\boldsymbol{\beta}$ selects only coefficients corresponding to the perturbable features to penalize. This trick thus inform us that the equivalence between robustness and regularization still stands with the prescence of non-perturbable features.

We end this section with a direct application in *LASSO*.

Corollary. *In LASSO, perturbing a subset of columns is equivalent to only penalizing the coefficients corresponding to the features in the same set. Namely, if* $\boldsymbol{D}$ *is defined as above,*

$$\min_{\boldsymbol{\beta}} \max_{||\boldsymbol{\Delta_i}||_2\leqslant\lambda} ||y - (X + \boldsymbol{\Delta D})\beta||_2 = \min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2 + \lambda \sum_{i=1}^{p-k} |\beta_i| \tag{3}$$

Proof. Set $z = y - X\boldsymbol{\beta}$, g be the l_2 norm, and h be the l_1 norm, we finish the proof. □

3 Applications in Robust Classification

The above technique can be applied to logistic regression, support vector machine, and optimal classification trees separately to extend its usage to dataset with non-perturbable features. We heavily use of the selection matrix $\boldsymbol{D}$ defined above to make the case.

Note that this section is inspired by and follows closely with (Bertsimas et al., 2019) Section 4. Throughout this section we consider a dataset $\{(x_i, y_i)\}_{i=1}^n$ with n observations and p features. For simplicity we limit our investigation to binary cases (i.e. $y_i \in \{-1, 1\}, \forall i \in \{1, ..., n\}$). We consider a specific uncertainty set $\mathcal{U}_q = \{\boldsymbol{\Delta} \in \mathbb{R}^{n\times p} |\ ||\boldsymbol{\Delta}_i||_q \leqslant \lambda, \forall i \in \{1, ..., n\}\}$.

In all cases, we can see that perturbing only a subset of features is equivalent to only penalizing the parameters for those features.

3.1 Dual Norm

Before showing the partially non-perturbable form of robust classification, we introduce a useful problem formulation that facilitate the following proofs.

Lemma. ***dual-norm problem*** *The dual-norm problem has the following form: for any $q \geqslant 1$, the l_q norm is denoted by $||x||_q$ for any $x \in \mathbb{R}^p$. For any $a \in \mathbb{R}^p$,*

$$\max_{||x||_q \leqslant \lambda} a^T x = \lambda ||a||_{q^*} \tag{4}$$

where l_{q^*} norm is the dual-norm of l_q norm, given by $q^* = \frac{1}{1-\frac{1}{q}}$.

The proof is completed by following the definition of dual norm.

3.2 Robust Soft-Margin Support Vector Machine

Robust SVM is given by

$$\min_{\beta,b} \max_{\boldsymbol{\Delta} \in \mathcal{U}_q} \sum_{i=1}^{n} \max(1 - y_i(\beta^T(x_i + \boldsymbol{\Delta}_i) - b), 0) \tag{5}$$

This can be modified such that only pertrubable features are selected for noisy perturbabtion by inserting the selection matrix $\boldsymbol{D}$:

$$\min_{\beta,b} \max_{\boldsymbol{\Delta} \in \mathcal{U}_q} \sum_{i=1}^{n} \max(1 - y_i(\beta^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) - b), 0) \tag{6}$$

and has the following equivalent regularization formulation:

Theorem. *The regularized counter part to Equation* (6) *is given by*

$$\begin{aligned} &\min_{\beta,b} \sum_{i=1}^{n} \xi_i \\ &s.t.\ \ y_i(\beta^T x_i - b) - \lambda ||\boldsymbol{D}\beta||_{q^*} \geqslant 1 - \xi_i \quad \forall i \in \{1, ..., n\} \\ &\qquad \xi_i \geqslant 0, \qquad\qquad\qquad\qquad\quad \forall i \in \{1, ..., n\} \end{aligned} \tag{7}$$

where l_{q^*} is the dual-norm of l_q

See Appendix B.1 for the complete proof.

3.3 Robust Logistic Regression

Robust logistic regression is formulated as

$$\max_{\beta,b} \min_{\boldsymbol{\Delta} \in \mathcal{U}_q} -\sum_{i=1}^{n} \log\left(1 + e^{-y_i(\beta^T(x_i + \boldsymbol{\Delta}_i) + b)}\right) \tag{8}$$

This can be similarly modified into partially perturbable version by inserting $\boldsymbol{D}$:

$$\max_{\beta,b} \min_{\boldsymbol{\Delta} \in \mathcal{U}_q} -\sum_{i=1}^{n} \log\left(1 + e^{-y_i(\beta^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) + b)}\right) \tag{9}$$

and has the following equivalent regularization formulation:

Theorem. *The regularized counterpart of Equation* (9) *is given by*

$$\max_{\beta,b} -\sum_{i=1}^{n} \log\left(1 + e^{-y_i(\beta^T x_i + b) + \lambda ||\boldsymbol{D}\beta||_{q^*}}\right) \tag{10}$$

where l_{q^*} *is the dual-norm of* l_q

See Appendix B.2 for the complete proof.

3.4 Optimal Classification Tree

Optimal classification tree has the following constraints:

$$\begin{aligned} a_j^T(x_i + \boldsymbol{\Delta}_i) + \epsilon \leqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^m \\ a_j^T(x_i + \boldsymbol{\Delta}_i) \geqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^u \end{aligned} \tag{11}$$

dictating the splits to the left and right respectively. Likewise this can be modified into non-pertrubable version:

$$\begin{aligned} a_j^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) + \epsilon \leqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^m \\ a_j^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) \geqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^u \end{aligned} \tag{12}$$

and has the following regularized counterpart:

Theorem. *The regularized counterpart of Equation* (12) *is given by*

$$\begin{aligned} a_j^T x_i + \lambda 1_{[j \leqslant p-k]} + \epsilon \leqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^m \\ a_j^T x_i + \lambda 1_{[j \leqslant p-k]} \geqslant b_j + M(1 - z_{im}), \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^u \end{aligned} \tag{13}$$

where $1_{[j \leqslant p-k]}$ *is 1 if feature* j *falls into the perturbable set and 0 otherwise.*

See Appendix B.3 for the complete proof.

4 Empirical Studies

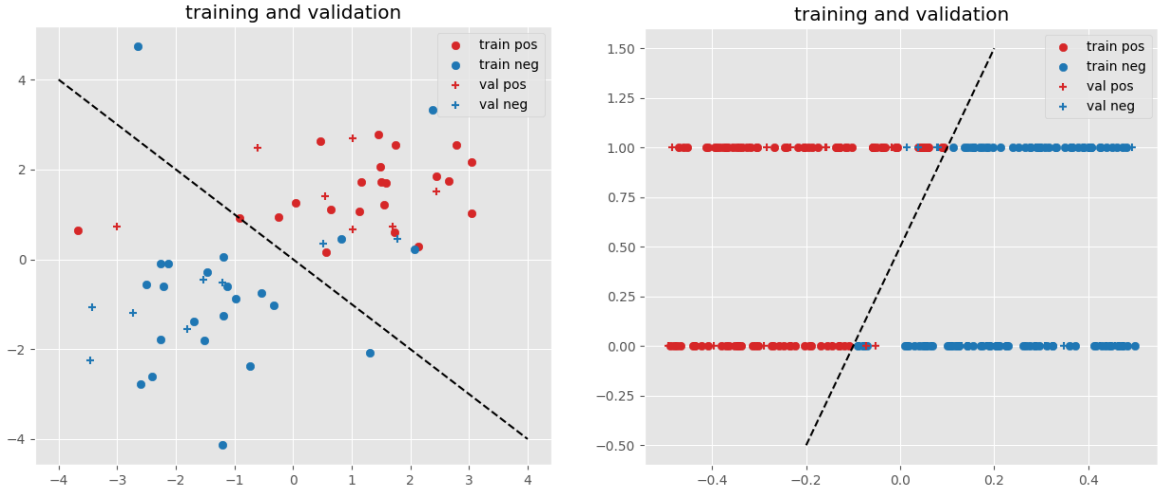
In this section, we demonstrate that with non-perturbable features, the robust formulation still beats the heuristic regularized version and the non-regularized naive version. We focus our investigation on robust svm and logistic regression with $q = q^* = 2$ (i.e. consider only the l_2 penalization in regularization and robustified version) by comparing the robustified version with the non-regularized version and naively regularized version. See section Appendix A for a discussion of the naively regularized problem formulation and how it differs from the robust-supported counterparts.

4.1 Synthetic Dataset

Data Generating Process We simulated two dataset both in $\mathbb{R}^2$: the first has both of its coordinate continuous while the second has one of the coordinate discrete (categorical).

The first dataset follows a Gaussian mixture with mean at $1.5\mathbf{1}$ and $-1.5\mathbf{1}$ with unit variance, where $\mathbf{1}$ is a vector of 1s. Noise are injected as another Gaussian with mean $\mathbf{0}$ and a variance of $3\mathbf{I}$. See more

details in (Bertsimas et al., 2019). The second dataset has its second coordinate categorical (e.g. either 1 or 0). The first coordiante are generated uniformly within the interval of $[-0.5, 0.5]$. Labels are assigned to 1 if $x_2 > 5x_1 + 0.5$ and -1 otherwise. Noise are generated in the same manner with labels randomly attached. The simulated dataset are partitioned into 75%/25% train-validation split to tune respective hyperparameters, and another much large number of testing dataset are generated in the same manner so that the training is non-trivial.



(a) both continuous, $\beta^{\text{true}} = [1, 1], b^{\text{true}} = 0$

(b) one categorical, $\beta^{\text{true}} = [-5, 1], b^{\text{true}} = 0.5$

Hyperparameters For regularized and robustified model, we consider $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$.

Performance To measure the performance of respective models, we compute the testing accuracy and testing deviation of trained boundary with the true boundary under 2000 different randomized inputs. The deviation in $\mathbb{R}^2$ is given as follows: suppose the ground-truth boundary is $\beta^{\text{true}} = [\beta_1^{\text{true}}, \beta_2^{\text{true}}]$ with intercept b^{true}, and the trained boundary is given by $\beta^{\text{train}} = [\beta_1^{\text{train}}, \beta_2^{\text{train}}]$ with intercept b^{train}, the deviation δ is measured by

$$\delta = \left\| \frac{\beta^{\text{true}}}{||\beta^{\text{true}}||_2} - \frac{\beta^{\text{train}}}{||\beta^{\text{train}}||_2} \right\|_2 + \left| \frac{b^{\text{true}}}{\beta_2^{\text{true}}} - \frac{b^{\text{train}}}{\beta_2^{\text{train}}} \right| \tag{14}$$

In particular, the former term measures the deviation in direction and the latter term measures the deviation in the second coordinate.

Note that in the first simulation, we may have three choices of perturbation: only pertube the first coordinate by considering the second without uncertainty, or only perturbe the second by treating the first without uncertainty, or both perturbable. The second data by design has only its first coordinate pertrubable. We thus report the performance in accuracy and deviation by model and perturbable terms in Table 1 and Table 2.

Observe that, overall, the partially perturbed model are not strong in terms of accuracy or deviaton and are inferior to the naively regularized counterparts and even to the unregularized models. Mostly only the fully perturbed model has performance gain. Two reasons may be at play: 1. the choices of hyperparameters λ are rather limited and thus the best performing one is not yet searched; 2. the simulated dataset is simple enough for unregularized model to perform well.

Table 1: Accuracy by Data and Selected Terms

data	terms	LR	LR Regularized	LR Robust	SVM	SVM Regularized	SVM Robust
continuous	both	0.9686 ± 0.015	0.9649 ± 0.023	**0.9705 ± 0.014**	0.9730 ± 0.014	**0.9755 ± 0.011**	0.9753 ± 0.012
	first	**0.9686 ± 0.015**	0.9657 ± 0.021	0.9642 ± 0.019	**0.9730 ± 0.014**	0.9693 ± 0.018	0.9692 ± 0.017
	second	**0.9686 ± 0.015**	0.9650 ± 0.023	0.9638 ± 0.019	**0.9730 ± 0.015**	0.9695 ± 0.018	0.9692 ± 0.018
categorical	first	**0.9425 ± 0.005**	0.9327 ± 0.018	0.9411 ± 0.006	**0.9348 ± 0.009**	0.9307 ± 0.013	0.9327 ± 0.010

Table 2: Deviation δ by Data and Selected Terms

data	terms	LR	LR Regularized	LR Robust	SVM	SVM Regularized	SVM Robust
continuous	both	1.4282 ± 17.4	1.3343 ± 5.88	**0.8093 ± 1.35**	0.5327 ± 0.87	0.4411 ± 0.51	**0.4406 ± 0.52**
	first	1.4282 ± 17.4	**0.8066 ± 1.00**	0.8386 ± 1.30	**0.5327 ± 0.87**	0.5460 ± 0.67	0.5398 ± 0.58
	second	1.4282 ± 17.4	**4.0780 ± 41**	10.4029 ± 255	**0.5327 ± 0.87**	21.8499 ± 192	16.8848 ± 143
categorical	first	**0.0491 ± 0.03**	0.1063 ± 0.16	0.0577 ± 0.04	1.0371 ± 0.09	1.0430 ± 0.16	**1.0365 ± 0.11**

The Consistency of λ One may suspect that partially perturbed dataset may lead to a different behavior in hyperparameter selection. It turns out that the bias in hyperparameter is quite consistent across different sets of perturbation. Figure 2 shows that in the first simulation the optimal λ picked by different randomization forms similar distributions. In particular, perturbing only one coordinate leads to similar λ as perturbing the full set.

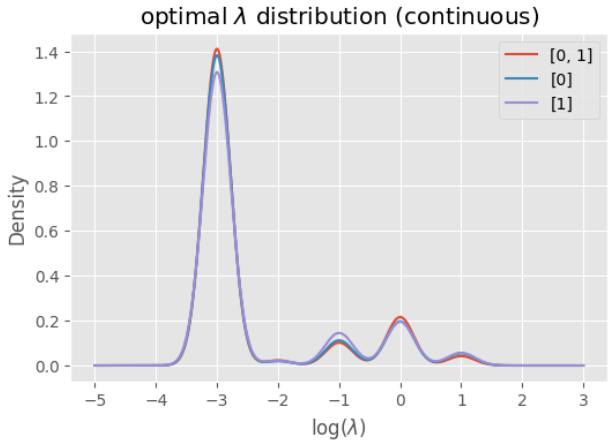


Figure 2: λ distribution for continuous simulated dataset with different set of perturbable terms

4.2 Real Dataset

Five real dataset are selected from UCI repository. Each has a mixture of numerical and categorical columns. See Table 3 for a short summary of the taken dataset. The preprocessing involves probabilistic imputations, One-Hot encoding of the categorical columns, and standardization of the numerical columns. Note that the intercept term is explicit in each of the problem formulation so that no intercept of 1s should be appended to the model matrix.

As reasoned above, we may only perturbe the numerical portion of the dataset (i.e. the number of predictors in the columns of **numerical (perturbable)** in Table 3). The performance is measured by a 5-fold cross validation and take the mean and standard deviation of accuracy by the best performing λ. See Table 4 for the full performance details.

Table 3: Summary of Dataset

dataset	observations	total predictors	categorical	numerical (perturbable)
australian	690	14	8	6
bands	425	17	6	11
heart	270	13	7	6
hepatitis	155	4	3	1
horse	300	20	13	7

Table 4: Real Dataset 5-fold Validation Accuracy

dataset	LR	LR Regularized	LR Robust	SVM	SVM Regularized	SVM Robust
australian	0.8246 ± 0.03	0.8260 ± 0.04	**0.8304 ± 0.04**	0.7478 ± 0.04	0.7319 ± 0.02	**0.7478 ± 0.04**
bands	0.6894 ± 0.07	**0.6894 ± 0.07**	0.6847 ± 0.07	0.6706 ± 0.02	0.6659 ± 0.02	**0.6706 ± 0.02**
heart	0.8370 ± 0.07	0.8370 ± 0.07	**0.8370 ± 0.07**	0.8259 ± 0.06	**0.8333 ± 0.05**	0.8259 ± 0.06
hepatitis	1.0000 ± 0.00	1.0000 ± 0.00	**1.0000 ± 0.00**	0.9226 ± 0.04	0.9226 ± 0.03	**0.9226 ± 0.04**
horse	0.7100 ± 0.05	0.7167 ± 0.05	**0.7200 ± 0.06**	0.6967 ± 0.04	0.7100 ± 0.06	**0.7167 ± 0.06**

Though failing in the simulated dataset, partially perturbed models perform better than their counterparts in real datasets, albeit not by a large margin. This demonstrates that the equivalence between robustness and regularization remains a powerful tool even with the prescence of non-perturbable features.

5 Future Works

Similar things can be shown for uncertainty in both the features and the label, following the rest of (Bertsimas et al., 2019). It is straightforward to extend the observations and proofs above to guarding against both labels and perturbable features.

In addition, it would be interesting to see if similar results hold for kernel SVM and its dual form formulation to generalie to training for hihgly non linearly separable datasets.

Finally, in certain cases we don't have the prior information that which sets of predictors don't have uncertainty. It may invole an "outer-optimization" problem to select the best sets of predictors to not perturbe. For example, MNIST hand-written digits are a collection of pictures of size 28 by 28 pixels. Not every pixel is uncertain. The boundary of images may not be subject to noisy perturbation by design. More generally, it would be interesting to develop a strategy to automatically select entries as non-pertrubable sets, although the interpretation issue would be non-trivial [1].

6 Code Availability

Code for this report is available at https://github.com/yangshengaa/robust_classification_partial. SVM is implemented through Gurobi and Logisitc Regression is written with PyTorch. We comment that the latter one is an unconstrained optimization problem and any standard automatic differentiation library can be employed.

[1] Idea from Kim

7 Conclusion

This reports discuss the equivalence between robustness and regularization under cases with non-perturbable features. The implication is that robustifying only the perturbable portion of the datset is equivalent to penalizing only the corresponding parameter estimates. We apply this characterization to robust classification and develop theorems for robust support vector machine, logistic regression, and optimal classification tree, and empirically verify the performance of the former two in both simulated and real dataset. Though the weak performance in simulated dataset motivates further studies in this topic, the slight improvement in real dataset suggests the power of equivalence between regularization and robustness over the naively regularized counterparts.

References

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Dimitris Bertsimas and Martin S Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942, 2018.

Dimitris Bertsimas and Jack Dunn. *Machine learning under a modern optimization lens*. Dynamic Ideas LLC, 2019.

Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

Apostolos Fertis. *A robust optimization approach to statistical estimation problems by Apostolos G. Fertis.* PhD thesis, Massachusetts Institute of Technology, 2009.

A Naively Regularized Formulation

These regularization are heuristic and may or may not coincide with the robustified version of regularization.

A.1 Regularized Support Vector Machine

The naively regularized method is given by

$$\begin{aligned} \min_{\beta,b} \;& n\lambda||\beta||_{q^*} + \sum_{i=1}^{n} \xi_i \\ s.t. \;& y_i(\beta^T x_i - b) \geqslant 1 - \xi_i \;\; \forall i \in \{1,...,n\} \\ & \xi_i \geqslant 0, \qquad\qquad\qquad \forall i \in \{1,...,n\} \end{aligned} \tag{15}$$

And the adaptation to partially perturbable version is straightforward by inserting the selection matrix $\boldsymbol{D}$ likewise.

It should be noted that the difference between Equation (15) and Equation (7) is small and under certain conditions (Fertis, 2009), the robust version is exactly the same as the heuristic regularized version.

A.2 Regularized Logistic Regression

The naively regularized logistic regression is given by

$$\max_{\beta,b} -\sum_{i=1}^{n} \log\left(1+e^{-y_i(\beta^T x_i+b)}\right) + n\lambda||\beta||_{q^*} \tag{16}$$

Pay close attention to the location of penalization. The exact equivalence between Equation (10) and Equation (16) can not be straightforwardly established as in the case of regularized support vector machine.

B Proof of the Equivalence with Non-perturbable Features in Robust Classification

In this section we show explicitly the proofs of the equivalence between robustness and regularization with non-perturbable features using the dual-norm problem results. Note that all proofs below closely follow those outlined in (Bertsimas et al., 2019).

B.1 Robust Support Vector Machine

Proof. Given Equation (6), we can reformulate the problem as the following by introducing a variable ξ_i:

$$\begin{aligned} &\min_{\beta,b} \sum_{i=1}^{n} \xi_i \\ &s.t.\ \ y_i(\beta^T(x_i+\boldsymbol{D}\boldsymbol{\Delta}_i)-b) \geqslant 1-\xi_i \quad \forall i \in \{1,...,n\} \\ &\quad \xi_i \geqslant 0, \qquad\qquad\qquad\qquad\quad \forall i \in \{1,...,n\} \end{aligned} \tag{17}$$

the first constraint can be rewritten as

$$y_i\beta^T\boldsymbol{D}\boldsymbol{\Delta}_i \geqslant 1-\xi_i - y_i(\beta^T x_i - b), \forall i \in \{1,...,n\}. \tag{18}$$

Since this holds for all i, we can restate the problem as

$$\min_{\boldsymbol{\Delta}\in\mathcal{U}_q} y_i\beta^T\boldsymbol{D}\boldsymbol{\Delta}_i \geqslant 1-\xi_i - y_i(\beta^T x_i - b), \forall i \in \{1,...,n\}. \tag{19}$$

Note that the solution to $\min_{\boldsymbol{\Delta}\in\mathcal{U}_q} y_i\beta^T\boldsymbol{D}\boldsymbol{\Delta}_i$ is $-\lambda||y_i\boldsymbol{D}\beta||_{q^*}|| = -||\boldsymbol{D}\beta||_{q^*}$, using the above dual-norm results and the knowledge that y_i is either 1 or -1. Plugging back we have Equation (7). □

B.2 Robust Logistic Regression

Proof. Note that $f(z) = -\log(1 + e^{-z})$ is a monotonically increasing function in z. This suggests that the inner optimization problem of Equation (9) is equivalent to

$$\min_{\boldsymbol{\Delta} \in \mathcal{U}_q} y_i(\beta^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) + b), \;\; \forall i \in \{1, ..., n\} \tag{20}$$

by making the connection of $z_i = y_i(\beta^T(x_i + \boldsymbol{D}\boldsymbol{\Delta}_i) + b)$. This in turn is equivalent to

$$\min_{\boldsymbol{\Delta} \in \mathcal{U}_q} \left(y_i(\beta^T x_i + b) - \max_{\boldsymbol{\Delta} \in \mathcal{U}_q} - y_i \beta^T \boldsymbol{D}\boldsymbol{\Delta}_i \right), \;\; \forall i \in \{1, ..., n\} \tag{21}$$

The inner maximization problem, but the dual-norm result, gives an optimal solution $\lambda ||\boldsymbol{D}\beta||_{q^*}$. Back-substituting z with this optimal value, we obtain the formulation of Equation (10) □

B.3 Optimal Classification Tree

Proof. Consider the partition for the first case: the first constraint in Equation (12) can be rewritten as

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_q} (a_j^T \boldsymbol{D}\Delta_i) \leqslant b_j + M(1 - z_{im}) - a_j^T x_i - \epsilon, \forall \boldsymbol{\Delta} \in \mathcal{U}_q, i \in \{1, ..., n\}, m \in \{1, ..., K\}, \forall j \in \mathcal{P}_l^m \tag{22}$$

Again, the first part maximization problem has a close-form solution given by $\lambda ||\boldsymbol{D}a_j||_{q^*}$. Now notice since $\boldsymbol{D}$ is 0 for row j when j is at the last k positions (i.e. positions for silencing non-perturbable parameters), and that a_j is everywhere 0 but 1 at one of its coordinate. Therefore this in aggregate is only one when $j \leqslant p - k$ (i.e. $\lambda ||\boldsymbol{D}a_j||_{q^*} = \lambda 1_{[j \leqslant p-k]}$).

□