# DATA-DEPENDENT GENERALIZATION PERFORMANCE ASSESSMENT VIA QUASICONVEX OPTIMIZATION

*Christopher P. Diehl and Ashley J. Llorens*

Applied Physics Laboratory
Johns Hopkins University
Laurel, Maryland 20723

## ABSTRACT

As compared to classical distribution-independent bounds based on the VC dimension, recent data-dependent bounds based on Rademacher complexity yield tighter upper bounds that may offer practical utility for model selection, as suggested by several investigations. We present an approach for kernel machine learning and generalization performance assessment that integrates concepts from prior work on Rademacher-type data-dependent generalization bounds and learning based on the optimization of quasiconvex losses. Our main contribution focuses on the direct estimation of the Rademacher penalty in order to obtain a tighter generalization bound. Specifically we define the optimization task for the case of learning with the ramp loss and show that direct estimation of the Rademacher penalty can be accomplished by solving a series of quadratic programming problems.

## 1. INTRODUCTION

When learning a classifier from data, our goal is to identify a function that provides good generalization performance on future examples drawn from the same random process that generated the training examples. Ideally we would like an approach that allows us to use the entire training set for learning and generalization performance assessment. The method would yield a classifier from the specified hypothesis class along with a rigorous, nontrivial bound on the generalization performance.

As compared to classical distribution-independent bounds based on the VC dimension, recent data-dependent bounds based on Rademacher complexity [1] yield tighter upper bounds that may offer practical utility for model selection, as suggested by several investigations [2, 3, 4, 5]. Furthermore they offer the potential to generate nontrivial upper bounds to assess classifier performance in an absolute sense which is beneficial in certain applications. To date these bounds have been mainly applied to fairly restricted learning problems where explicit training error minimization is possible [2, 3, 4]. [5] introduces a method that supports joint learning and feature selection for linear models through the direct optimization of a Rademacher-type bound.

In this paper, we present an approach for kernel machine learning and generalization performance assessment that integrates concepts from prior work on Rademacher-type data-dependent generalization bounds and learning based on the optimization of quasiconvex losses. Our main contribution focuses on the direct estimation of the Rademacher penalty [4] in order to obtain a tighter generalization bound. Specifically we define the optimization task for the case of learning with the ramp loss and show that direct estimation of the Rademacher penalty can be accomplished by solving a series of quadratic programming problems.

## 2. THE LEARNING TASK

We will focus on the task of learning a binary classifier from a labeled training sample $\mathcal{S} = \{(x_i, y_i) \mid i \in \{1, ..., n\}, x_i \in \mathbb{R}^m, y_i \in \{\pm 1\}\}$ that is generated from an independent, identically distributed (IID) random process. Let $\phi(h(x), y)$ be the loss function of interest. The *true risk* $L(h)$ for a classifier $h$ is $L(h) = \mathbf{E}\phi(h(x), y)$. Given a hypothesis class $\mathcal{H} = \{h\}$, the classifier $\hat{h}$ is selected from $\mathcal{H}$ that minimizes the empirical estimate $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} \phi(h(x_i), y_i)$ of the risk over the training data. This induction principle is known as *empirical risk minimization (ERM)*.

In the limit of infinite training data, we desire the empirical risk to converge to the true risk uniformly across the hypothesis class. Furthermore, we desire the deviation between the empirical risk and the true risk to be reasonably small for a finite sample so that we have confidence in the generalization performance of the resulting classifier. We will examine data-dependent *uniform deviation bounds* that place a bound on the maximum deviation between the empirical and true risk over the hypothesis class $\sup_{h \in \mathcal{H}} (L(h) - L_n(h))$. Given $L(\hat{h}) \leq L_n(\hat{h}) +$

$\sup_{h\in\mathcal{H}}(L(h)-L_n(h))$, controlling the maximum deviation constrains the upper bound on the true risk.

## 3. CONCENTRATION

The approach we will review to bounding the deviation between the empirical and true risk relies on the idea of concentration of a random variable. A concentrated random variable is one which is very likely to assume values close to its expectation. A result applied repeatedly is McDiarmid's inequality which allows one to bound the probability that a function of independent variables deviates from its expectation greater than a specified level. Specifically, let $x_1, x_2, \ldots, x_n$ be independent random variables taking values in a set $A$. Assume $g : A^n \to \mathbb{R}$ satisfies

$$\sup_{X_1,\ldots,X_n,\hat{X}_i\in A} \left|g(X_1,\ldots,X_n)-g(X_1,\ldots,\hat{X}_i,X_{i+1},\ldots,X_n)\right|\leq c_i \quad (1)$$

for $1 \leq i \leq n$.

Then for all $\epsilon > 0$,

$$\mathrm{P}(g(x_1,\ldots,x_n)-\mathbf{E}g(x_1,\ldots,x_n)\geq\epsilon)\leq\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}c_i^2}\right) \quad (2)$$

and

$$\mathrm{P}(\mathbf{E}g(x_1,\ldots,x_n)-g(x_1,\ldots,x_n)\geq\epsilon)\leq\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}c_i^2}\right) \quad (3)$$

[6].

Consider the maximum deviation between the empirical and true risk once again

$$\sup_{h\in\mathcal{H}}(L(h)-L_n(h)) = \sup_{h\in\mathcal{H}}\left(\mathbf{E}\phi(yh(x))-\frac{1}{n}\sum_{i=1}^{n}\phi(y_ih(x_i))\right). \quad (4)$$

Let us define $z_i = \phi(y_ih(x_i))$ and

$$g(z_1,\ldots,z_n) = \sup_{h\in\mathcal{H}}\left(\mathbf{E}z_1-\frac{1}{n}\sum_{i=1}^{n}z_i\right). \quad (5)$$

In order for this function of the training sample to be concentrated, the maximum deviation in $z_i$ must be bounded, either due to a constraint on $\phi$ or $h$. Let us assume $0 \leq z_i \leq B$ which implies $c_i = \frac{B}{n}$. If (2) holds with probability at most $\delta$, then we find with probability at least $1-\delta$ [7],

$$\sup_{h\in\mathcal{H}}(L(h)-L_n(h)) \leq \mathbf{E}\sup_{h\in\mathcal{H}}(L(h)-L_n(h))+B\sqrt{\frac{-\log\delta}{2n}}. \quad (6)$$

## 4. SYMMETRIZATION

The next objective is to bound the expectation of the deviation. This is achieved through the introduction of a *ghost sample* $S'$ of size $n$ from the same underlying distribution. The expected value of the maximum deviation between the true risk and the empirical risk is then bounded by the expected value of the maximum deviation between the empirical risks on the two samples $S$ and $S'$ through the application of Jensen's inequality yielding [7]

$$\begin{aligned}&\mathbf{E}\sup_{h\in\mathcal{H}}(L(h)-L_n(h))\\ &\quad\leq \mathbf{E}_{SS'}\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i'h(x_i'))-\phi(y_ih(x_i))\right). \quad (7)\end{aligned}$$

At this point, a series of independent and identically distributed (IID) *Rademacher random variables* $\{\sigma_i \in \{\pm 1\} : \forall\, i \in \{1,\ldots,n\}\}$ with uniform distributions are introduced to equation 7. These variables represent the introduction of random exchanges of examples between the two samples. This ultimately yields the following bound [7].

$$\begin{aligned}&\mathbf{E}\sup_{h\in\mathcal{H}}(L(h)-L_n(h))\\ &\quad\leq \mathbf{E}_{SS'\sigma}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[\phi(y_i'h(x_i'))-\phi(y_ih(x_i))\right]\\ &\quad\leq R(\phi\circ\mathcal{F}) = \mathbf{E}_{S\sigma}\sup_{h\in\mathcal{H}}\frac{2}{n}\sum_{i=1}^{n}\sigma_i\phi(y_ih(x_i)) \quad (8)\end{aligned}$$

The *Rademacher complexity* $R(\phi\circ\mathcal{F})$, where $\mathcal{F} = \{yh(x) : \forall\; h \in \mathcal{H}\}$, provides a measure of the richness of the hypothesis class $\mathcal{H}$ in terms of the expected value of the maximum deviation achievable in the empirical risks over random partitions of realizations of the training set. A realization of the Rademacher random variables partitions a training set into two approximately equal-sized subsamples with high probability. Naturally if the hypothesis class is very general, there is increased risk of overfitting, which can manifest itself in terms of a significant deviation in performance across the two subsamples. Our aim is to identify a hypothesis class that allows us to achieve an acceptably low risk without excessive complexity. The Rademacher complexity is an appealing measure because it allows for a completely data-dependent assessment of generalization performance.

## 5. DATA-DEPENDENT GENERALIZATION PERFORMANCE ASSESSMENT

With the results introduced so far, we can state the following bound on the generalization performance of the classifier.

With probability at least $1 - \delta$,

$$L(h) \leq L_n(h) + R(\phi \circ \mathcal{F}) + B\sqrt{\frac{-\log \delta}{2n}}. \qquad (9)$$

This bound is still unappealing due to the expectation over the sample and the Rademacher random variables in the Rademacher complexity term. Thankfully we can avoid this complication by appealing once again to concentration and McDiarmid's inequality. We examine two bounds that enable different paths towards a computable bound. Their respective roles will be made clear in the next section.

The *empirical Rademacher complexity* is defined as

$$R_n(\phi \circ \mathcal{F}) = \mathbf{E}_\sigma \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \phi(y_i h(x_i)), \qquad (10)$$

where $z_i = \phi(y_i h(x_i))$ and

$$g(z_1, ..., z_n) = \mathbf{E}_\sigma \sup_{h \in \mathcal{H}} \left( \frac{2}{n} \sum_{i=1}^{n} \sigma_i z_i \right). \qquad (11)$$

$R_n(\phi \circ \mathcal{F})$ varies by at most $\frac{2B}{n}$ since $0 \leq z_i \leq B$; therefore $c_i = \frac{2B}{n}$. If (3) holds with probability at most $\delta$, then with probability at least $1 - \delta$,

$$R(\phi \circ \mathcal{F}) \leq R_n(\phi \circ \mathcal{F}) + 2B\sqrt{\frac{-\log \delta}{2n}}. \qquad (12)$$

Similarly, the *Rademacher penalty* is defined as

$$R_n(\phi \circ \mathcal{F} \mid \sigma) = \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \phi(y_i h(x_i)), \qquad (13)$$

where $z_i = \sigma_i \phi(y_i h(x_i))$ and

$$g(z_1, ..., z_n) = \sup_{h \in \mathcal{H}} \left( \frac{2}{n} \sum_{i=1}^{n} z_i \right).$$

$R_n(\phi \circ \mathcal{F} \mid \sigma)$ varies by at most $\frac{4B}{n}$ since $-B \leq z_i \leq B$; therefore $c_i = \frac{4B}{n}$. If (3) holds with probability at most $\delta$, then with probability at least $1 - \delta$,

$$R(\phi \circ \mathcal{F}) \leq R_n(\phi \circ \mathcal{F} \mid \sigma) + 4B\sqrt{\frac{-\log \delta}{2n}}. \qquad (14)$$

We can now assert that if (9) and (12) both hold with probability at least $1 - \frac{\delta}{2}$, then through the union bound

$$L(h) \leq L_n(h) + R_n(\phi \circ \mathcal{F}) + 3B\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \qquad (15)$$

holds with probability at least $1 - \delta$. Similarly if (9) and (14) both hold with probability at least $1 - \frac{\delta}{2}$, then

$$L(h) \leq L_n(h) + R_n(\phi \circ \mathcal{F} \mid \sigma) + 5B\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \qquad (16)$$

holds with probability at least $1 - \delta$. This bound holds for a single realization of the sample and the Rademacher random variables.

## 6. BOUNDING THE EMPIRICAL RADEMACHER COMPLEXITY

In scenarios where it is computationally intractable to directly compute the Rademacher complexity term in (15), we can still construct a data-dependent generalization bound from (15) by bounding the empirical Rademacher complexity. It can be shown that for a bounded loss $\phi$ with Lipschitz coefficient $\mathcal{L}_\phi$, the true risk is bounded by

$$L(h|\phi) \leq L_n(h|\phi) + 2\mathcal{L}_\phi R_n(\mathcal{F}) + 3B\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \qquad (17)$$

with probability at least $1 - \delta$. The remaining task is to bound $R_n(\mathcal{F})$.

At this point, we will assume that the hypothesis class is $\mathcal{H} = \{h(x) = w^{\mathrm{T}}\Phi(x) : \forall\ ||w|| \leq W\}$, a set of generalized linear classifiers with weight vectors of bounded norm. This leads to the following upper bound

$$R_n(\mathcal{F}) \leq \frac{2W}{n} \left( \sum_{i=1}^{n} K(x_i, x_i) \right)^{\frac{1}{2}} \qquad (18)$$

where the kernel function $K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}}\Phi(x_j)$ [8]. Substituting this result into (17) yields the bound

$$L(h|\phi) \leq L_n(h|\phi) + \frac{4\mathcal{L}_\phi W}{n} \left( \sum_{i=1}^{n} K(x_i, x_i) \right)^{\frac{1}{2}} + 3B\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \qquad (19)$$

that holds with probability at least $1 - \delta$.

It is worth pausing for a moment to reflect on the implications of this bound. First note that this bound allows one to make statements about the generalization performance of any classifier $h \in \mathcal{H}$ with respect to loss $\phi$, even if $h$ is chosen through the minimization of a different loss $\phi'$. Clearly if we are optimizing with respect to $\phi'$, the empirical risk term $L_n(h|\phi)$ will not be minimized in the bound; yet we can still define a bound.

In the derivation of the bound, two constraints were introduced on the loss $\phi$; namely that $\phi$ has a finite Lipschitz coefficient $\mathcal{L}_\phi$ and $\phi(yh(x))$ is bounded. For simplicity, let us consider bounded losses where $0 \leq \phi(x) \leq B$. For example, if we desire a bound on the classification error rate, the standard approach is to specify a loss that upper bounds the classification error loss $\mathcal{I}_{yh(x)<0}$ yet satisfies the Lipschitz continuity constraint. One such loss is the ramp loss (or clipped hinge loss) where

$$\phi_{r|s}(x) = \begin{cases} 1 & \text{if } x < 0 \\ 1 - \frac{x}{s} & \text{if } 0 \leq x \leq s \\ 0 & \text{otherwise} \end{cases} \qquad (20)$$

with Lipschitz coefficient $\mathcal{L}_{\phi_{r|s}} = \frac{1}{s}$ [8]. Using this loss, we can specify the following bound on the classification error rate

$$\begin{aligned} L(h|\mathcal{I}_{x<0}) &\le L(h|\phi_{r|s}) \\ &\le L_n(h|\phi_{r|s}) + \frac{4W}{sn}\left(\sum_{i=1}^{n} K(x_i,x_i)\right)^{\frac{1}{2}} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \end{aligned} \quad (21)$$

with probability at least $1-\delta$.

While the bound introduced in (21) appears favorable, there is still room for improvement. Many practical learning algorithms involve the optimization of a convex loss that is by definition not bounded. Therefore, as highlighted earlier, the empirical risk term in the bound will not be minimized for a classifier derived from optimization of a surrogate loss. Ideally we would like to minimize this term through direct optimization of the bounded loss.

The more significant source of slack in (21) comes from the upper bound on the empirical Rademacher complexity. The derivation leading to (18) employs two bounds for the hypothesis class of norm-constrained generalized linear models to address the expectation over the Rademacher random variables and the supremum over the hypothesis class. This introduces the linear dependence on the weight vector norm bound $W$ which is particularly concerning. As the difficulty of the classification problem increases, necessitating a larger weight vector norm, the bound can grow arbitrarily large and become dominated by this linear dependence.

Instead of bounding the empirical Rademacher complexity, we will now explore an alternative approach that involves directly computing the Rademacher penalty for a specific realization of the Rademacher random variables. In this scenario, we will shift our attention from bound (15) to (16) as our starting point.

## 7. A TIGHTER BOUND THROUGH OPTIMIZATION

The data-dependent generalization bound based on the Rademacher penalty offers the potential to pose learning algorithms with tighter bounds that are directly evaluated through optimization. To achieve this result, we are trading convexity in the loss to obtain a tighter bound on generalization performance. In this section, we examine the details of the optimization problems associated with a bounded, quasiconvex loss.

For a bounded loss $\phi$ that upper bounds the classification error, we can state the following upper bound on the true classification error

$$\begin{aligned} L(h|\mathcal{I}_{x<0}) &\le L(h|\phi) \\ &\le L_n(h|\phi) + R_n(\phi\circ\mathcal{F}\mid\sigma) + 5B\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \end{aligned} \quad (22)$$

with probability at least $1-\delta$ using (16). During learning, the objective is to identify a classifier that minimizes the bound by searching over a series of nested hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_m$. As mentioned earlier, our interest is in learning kernel machines of the form $h(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ where $K(x,y) = \Phi(x)^{\mathsf{T}}\Phi(y)$. The nested hypothesis classes are increasingly larger sets of kernel machines with weight vectors satisfying a norm bound. Let $k_{ij} = K(x_i, x_j)$, $k_i = [k_{1i}, k_{2i}, \ldots, k_{ni}]^{\mathsf{T}}$ and $K = [k_1, k_2, \ldots, k_n]$ and $\mathcal{H} = \{h : \alpha^{\mathsf{T}}K\alpha \le \bar{W}\}$. The hypothesis classes are therefore indexed by their corresponding weight vector norm bounds $\bar{W}_1 \le \bar{W}_2 \le \ldots \le \bar{W}_m$.

For a given hypothesis class with norm bound $\bar{W}$, we must address two optimization problems to select a classifier and obtain a generalization bound. The first step is to estimate the classifier parameters $\alpha_*$ minimizing the empirical risk

$$\alpha_*(\bar{W}) = \arg \min_{\substack{\alpha \\ \alpha^{\mathsf{T}}K\alpha \le \bar{W}}} L_n(h|\phi) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i k_i^{\mathsf{T}}\alpha). \quad (23)$$

The second step is to estimate the Rademacher penalty associated with the hypothesis class for the specific realization of the Rademacher random variables.

$$R_n(\phi\circ\mathcal{F}\mid\sigma,\bar{W}) = \max_{\substack{\alpha \\ \alpha^{\mathsf{T}}K\alpha \le \bar{W}}} \left(\frac{2}{n}\sum_{\sigma_i=1}\phi(y_i k_i^{\mathsf{T}}\alpha) - \frac{2}{n}\sum_{\sigma_j=-1}\phi(y_j k_j^{\mathsf{T}}\alpha)\right). \quad (24)$$

With these parameters identified, the resulting generalization bound for the classifier $h_*$ is

$$\begin{aligned} L(h|\mathcal{I}_{x<0}) \le \frac{1}{n}\sum_{i=1}^{n}\phi(y_i k_i^{\mathsf{T}}\alpha_*(\bar{W})) \\ + R_n(\phi\circ\mathcal{F}\mid\sigma,\bar{W}) + 5B\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \end{aligned} \quad (25)$$

with probability at least $1-\delta$.

Since we are interested in a subset of the solutions along the trajectory parameterized by $\bar{W}$, we address a corresponding set of unconstrained optimization problems to identify the classifiers and estimate the Rademacher penalties. The unconstrained minimization problem to solve for empirical risk minimization is

$$\alpha_*(\lambda) = \arg\min_{\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i k_i^{\mathsf{T}}\alpha) + \lambda\alpha^{\mathsf{T}}K\alpha\right). \quad (26)$$

For a given $\lambda$, there exists a corresponding $\bar{W}(\lambda) = \alpha_*(\lambda)^{\mathsf{T}}K\alpha_*(\lambda)$ for which $\lambda$ is the resulting Lagrange multiplier in the constrained optimization problem in (23). This implies that $\lambda$ provides an alternate parameterization of the trajectory of solutions that we can trace without the burden of constraints.

The unconstrained maximization problem to solve in order to compute the Rademacher penalties for the resulting classifiers is

$$\alpha'_*(\beta) = \arg\max_\alpha \left( \frac{2}{n} \sum_{\sigma_i=1} \phi(y_i k_i^\mathrm{T} \alpha) \right. \\ \left. - \frac{2}{n} \sum_{\sigma_j=-1} \phi(y_j k_j^\mathrm{T} \alpha) - \beta \alpha^\mathrm{T} K \alpha \right) . \quad (27)$$

As with the unconstrained minimization problem, $\beta$ represents the Lagrange multiplier for the corresponding constrained optimization problem in (24) with

$$\bar{W}(\beta) = \alpha'_*(\beta)^\mathrm{T} K \alpha'_*(\beta).$$

By solving the maximization problem for a series of values $\{\beta_i\}$, we obtain samples $\{ \big(\bar{W}(\beta_i), R_n(\phi \circ \mathcal{F} \mid \sigma, \bar{W}(\beta_i))\big)\}$ on the Rademacher penalty curve as a function of $\bar{W}$ supporting estimation of the needed Rademacher penalties for $\{\bar{W}(\lambda_i)\}$ through interpolation.

## 8. QUASICONVEX OPTIMIZATION VIA CCCP

For both empirical risk minimization and computation of the Rademacher penalty, the core task is to minimize a quasiconvex function $g$ that can be expressed as a sum of a convex function $g_v$ and a concave function $g_c$. The approach we will employ to minimize functions of this form is the ConCave-Convex Procedure (CCCP) [9]. CCCP identifies a local minimum by solving a sequence of convex minimization problems where the concave function is majorized by a first-order approximation about the current parameters. After each convex minimization problem is solved, the first-order approximation is recomputed and the process continues until a minimum is reached.

For empirical risk minimization, the goal is to minimize

$$\min_\alpha \frac{1}{n} \sum_{i=1}^{n} \phi_v(y_i k_i^\mathrm{T} \alpha) + \phi_c(y_i k_i^\mathrm{T} \alpha) + \lambda \alpha^\mathrm{T} K \alpha \quad (28)$$

where $\phi_v$ and $\phi_c$ are the convex and concave components of $\phi$ respectively. Beginning with $\alpha_0 = 0$, the convex surrogate

$$\min_{\alpha_{k+1}} \frac{1}{n} \sum_{i=1}^{n} \phi_v(y_i k_i^\mathrm{T} \alpha_{k+1}) + \\ \nabla_\alpha \phi_c(y_i k_i^\mathrm{T} \alpha)|_{\alpha=\alpha_k}^\mathrm{T} \alpha_{k+1} + \lambda \alpha_{k+1}^\mathrm{T} K \alpha_{k+1}. \quad (29)$$

is minimized during each iteration, yielding the parameters $\alpha_{k+1}$ from the convex approximation centered about the parameters $\alpha_k$ from the previous iteration.

We proceed similarly with computing the Rademacher penalty. The objective in this case is to maximize

$$\max_\alpha \frac{2}{n} \sum_{\sigma_i=1} \phi(y_i k_i^\mathrm{T} \alpha) \\ - \frac{2}{n} \sum_{\sigma_j=-1} \phi(y_j k_j^\mathrm{T} \alpha) - \beta \alpha^\mathrm{T} K \alpha. \quad (30)$$

Grouping convex and concave terms yields

$$\max_\alpha \overbrace{\frac{2}{n} \sum_{\sigma_i=1} \phi_v(y_i k_i^\mathrm{T} \alpha) - \frac{2}{n} \sum_{\sigma_j=-1} \phi_c(y_j k_j^\mathrm{T} \alpha)}^{g_v(\alpha)} + \\ \underbrace{\frac{2}{n} \sum_{\sigma_i=1} \phi_c(y_i k_i^\mathrm{T} \alpha) - \frac{2}{n} \sum_{\sigma_j=-1} \phi_v(y_j k_j^\mathrm{T} \alpha) - \beta \alpha^\mathrm{T} K \alpha}_{g_c(\alpha)} \quad (31)$$

$$= \max_\alpha g_v(\alpha) + g_c(\alpha). \quad (32)$$

Beginning with $\alpha_0 = 0$, the concave surrogate

$$\max_{\alpha_{k+1}} g_c(\alpha_{k+1}) + \nabla_\alpha g_v(\alpha)|_{\alpha=\alpha_k}^\mathrm{T} \alpha_{k+1} \quad (33)$$

is maximized during each iteration, yielding the parameters $\alpha_{k+1}$ from the concave approximation centered about the parameters $\alpha_k$ from the previous iteration.

We now consider the specific problems induced when $\phi(x)$ is a generalized ramp loss

$$\phi(x) = \begin{cases} 1 - \gamma x & \text{if } x < 0 \\ 1 - x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (34)$$

where $0 < \gamma \leq 1$. In [10], scalable learning of kernel machines optimized for the ramp loss was addressed through the application of CCCP as outlined above. Therefore we focus on the problem of computing the Rademacher penalty in this scenario.

The generalized ramp loss can be decomposed into a difference of shifted hinge losses

$$\begin{aligned} \phi(x) &= \phi_v(x) + \phi_c(x) \\ &= [1 - x]_+ - [-(1-\gamma)x]_+. \end{aligned} \quad (35)$$

For $\gamma > 0$, the generalized ramp loss is strictly quasiconvex which guarantees a unique solution [11]. By choosing a small value for $\gamma$ initially or annealing toward 0, we can capitalize on the strict quasiconvexity of the loss.

Substituting in for $\phi_c$ and $\phi_v$, we obtain the following for the concave and convex components

$$\begin{aligned} g_c(\alpha) &= -\frac{2}{n} \sum_{i,\sigma_i=1} [-(1-\gamma) y_i k_i^\mathrm{T} \alpha]_+ \\ &\quad -\frac{2}{n} \sum_{j,\sigma_j=-1} [1 - y_j k_j^\mathrm{T} \alpha]_+ - \beta \alpha^\mathrm{T} K \alpha \end{aligned} \quad (36)$$

$$
\begin{aligned}
g_v(\alpha) \quad &= \quad \frac{2}{n} \sum_{i,\sigma_i=1} [1 - y_i k_i^{\mathrm{T}} \alpha]_+ \\
&\quad + \frac{2}{n} \sum_{j,\sigma_j=-1} [-(1-\gamma) y_j k_j^{\mathrm{T}} \alpha]_+ . \qquad (37)
\end{aligned}
$$

We will define the subderivative of the hinge loss $[\cdot]_+$ as

$$
\frac{\partial}{\partial x}[x]_+ = \left\{ \begin{array}{ll} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{array} \right. . \qquad (38)
$$

The subgradient $\nabla_\alpha g_v(\alpha)$ is then

$$
\begin{aligned}
\nabla_\alpha g_v(\alpha) \quad &= \quad -\frac{2}{n} \sum_{i,\sigma_i=1} y_i k_i p(\alpha)_i \\
&\quad -\frac{2}{n} \sum_{j,\sigma_j=-1} (1-\gamma) y_j k_j q(\alpha,\gamma)_j \quad (39)
\end{aligned}
$$

where $p(\alpha)_i = \mathcal{I}_{y_i k_i^{\mathrm{T}} \alpha < 1}$ and $q(\alpha,\gamma)_i = \mathcal{I}_{(1-\gamma) y_i k_i^{\mathrm{T}} \alpha < 0}$. By introducing slack variables for the hinge losses, we obtain the following equivalent quadratic program for (33)

$$
\begin{array}{cc}
\min\limits_{\alpha,\epsilon} & \beta \alpha^{\mathrm{T}} K \alpha - \nabla_\alpha g_v(\alpha)|_{\alpha=\alpha_0}^{\mathrm{T}} \alpha + \frac{2}{n} \sum_i \epsilon_i \\
\text{subject to} & (1-\gamma) y_i k_i^{\mathrm{T}} \alpha \geq -\epsilon_i,\, \sigma_i = 1 \\
& y_i k_i^{\mathrm{T}} \alpha \geq 1 - \epsilon_i,\, \sigma_i = -1 \\
& \epsilon_i \geq 0 \ \ \forall \, i \in \{1,\ldots,n\}.
\end{array}
$$

We omit the derivation of the dual optimization problem due to space constraints. Given the range of scalable approaches for solving quadratic programs, we expect that direct computation of the Rademacher penalty is feasible with runtime complexity similar to that of the learning with the ramp loss.

## 9. DISCUSSION AND CONCLUSION

With the machinery introduced to compute the Rademacher penalty directly, let us consider what claims can be made about the relative performance of the bounds (15) and (16). Consider the case where the ramp loss is employed along with normalized kernels where $K(x,x) = 1$. If the Rademacher penalty-based bound (16) outperforms (15),

$$
R_n(\phi \circ \mathcal{F} \mid \sigma) \leq \frac{4W}{s\sqrt{n}} - \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}. \qquad (40)
$$

Since $0 \leq R_n(\phi \circ \mathcal{F} \mid \sigma) \leq \frac{2n_+}{n}$ where $n_+$ is the number of positive realizations of the Rademacher random variables, we can identify two transition points. When $W \leq s\sqrt{\log \frac{2}{\delta}}$, the above condition is violated and the Rademacher penalty-based bound underperforms. When $W \geq s\sqrt{\log \frac{2}{\delta}} + \frac{sn_+}{2\sqrt{n}}$, the bound outperforms. In the intervening range, no claims can be made.

These transition points suggest the empirical Rademacher complexity-based bound (15) may outperform the Rademacher penalty-based bound (16) in small sample scenarios while (16) may excel in challenging tasks where more functional complexity is required. One may find also that even with direct computation of the Rademacher penalty, this global measure of complexity is still insufficient, motivating the need for local Rademacher averages [12] that reflect the fact that classifiers are chosen from a small set with low empirical risk. Yet as earlier work on decision tree pruning [4] and feature selection [5] suggests, the use of Rademacher-based bounds such as (15) and (16) provides practical performance gains. Our future work will be focused on conducting extensive computational studies to understand the tradeoffs and benefits provided by direct estimation of the Rademacher penalty for kernel machine learning.

### A. REFERENCES

[1] Vladimir Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, July 2001.

[2] Fernando Lozano, "Model selection using Rademacher penalization," in *Proceedings of the Second ICSC Symposia on Neural Computation*. 2000, ICSC Academic Press.

[3] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi, "Model selection and error estimation," *Machine Learning*, vol. 48, pp. 85–113, 2002.

[4] Matti Kääriäinen, Tuomo Malinen, and Tapio Elomaa, "Selective Rademacher penalization and reduced error pruning of decision trees," *Journal of Machine Learning Research*, vol. 5, pp. 1107–1126, 2004.

[5] Dori Peleg and Ron Meir, "A feature selection algorithm based on the global minimization of a generalization error bound," in *Advances in Neural Information Processing Systems (NIPS) 18*, 2004.

[6] Colin McDiarmid, "On the method of bounded differences," in *Survey of Combinatorics*, pp. 148–188. Cambridge University Press, 1989.

[7] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi, "Theory of classification: A survey of recent advances," *ESAIM: Probability and Statistics*, 2005.

[8] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[9] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (cccp)," in *Advances in Neural Information Processing Systems 14*. 2002, MIT Press.

[10] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou, "Trading convexity for scalability," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[11] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[12] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson, "Local rademacher complexities," *The Annals of Statistics*, vol. 33, pp. 1497–1537, 2005.