# Smoothing Multivariate Performance Measures

**Xinhua Zhang**
Department of Computing Science
University of Alberta
Alberta, T6G 2E8, Canada
xinhua2@ualberta.ca

**Ankan Saha**
Department of Computer Science
University of Chicago
Chicago, IL 60637, USA
ankans@cs.uchicago.edu

**S.V. N. Vishwanathan**
Department of Statistics and
Department of Computer Science
Purdue University, IN 47907, USA
vishy@stat.purdue.edu

## Abstract

A Support Vector Method for multivariate performance measures was recently introduced by Joachims (2005). The underlying optimization problem is currently solved using cutting plane methods such as SVM-Perf and BMRM. One can show that these algorithms converge to an $\epsilon$ accurate solution in $O\left(\frac{1}{\lambda\epsilon}\right)$ iterations, where $\lambda$ is the trade-off parameter between the regularizer and the loss function. We present a smoothing strategy for multivariate performance scores, in particular precision/recall break-even point and ROCArea. When combined with Nesterov's accelerated gradient algorithm our smoothing strategy yields an optimization algorithm which converges to an $\epsilon$ accurate solution in $O^*\left(\min\left\{\frac{1}{\epsilon}, \frac{1}{\sqrt{\lambda\epsilon}}\right\}\right)$ iterations. Furthermore, the cost per iteration of our scheme is the same as that of SVM-Perf and BMRM. Empirical evaluation on a number of publicly available datasets shows that our method converges significantly faster than cutting plane methods without sacrificing generalization ability.

## 1 Background and Introduction

Different kinds of applications served by machine learning algorithms have varied and specific measures to judge the performance of the algorithms. In this paper we focus on efficient algorithms for directly optimizing multivariate performance measures such as precision/recall break-even point (PRBEP) and area under the Receiver Operating Characteristic curve (ROCArea). Given a training set with $n$ examples $\mathcal{X} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{+1, -1\}$, Joachims (2005) proposed an elegant formulation for this problem which minimizes the following regularized risk:

$$\min_{\mathbf{w}} \; J(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + R_{\text{emp}}(\mathbf{w}). \tag{1}$$

Here $\frac{1}{2}\|\mathbf{w}\|^2$ is the regularizer, $\lambda > 0$ is a trade-off parameter and the empirical risk $R_{\text{emp}}$ for contingency table based multivariate performance measures is

$$R_{\text{emp}}(\mathbf{w}) = \max_{\mathbf{z}\in\{-1,1\}^n}\left[\Delta(\mathbf{z},\mathbf{y}) + \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i\rangle (z_i - y_i)\right]. \tag{2}$$

Here, $\Delta(\mathbf{z}, \mathbf{y})$ denotes the multivariate discrepancy between the correct labels $\mathbf{y} := (y_1, \ldots, y_n)^\top$ and a candidate labeling $\mathbf{z}$ (Joachims, 2005), and $\langle \cdot, \cdot \rangle$ denotes the Euclidean dot product. In order to compute the multivariate discrepancy for the PRBEP, which is the main focus of our work, we need the false positive and false negative rates, which are defined as

$$b = \sum_{i\in\mathcal{P}} \delta(z_i = -1), \text{ and } c = \sum_{j\in\mathcal{N}} \delta(z_j = 1).$$

Here $\delta(x) = 1$ if $x$ is true and 0 otherwise, while $\mathcal{P}$ and $\mathcal{N}$ denote the set of indices of positive ($y_i = +1$) and negative ($y_i = -1$) examples respectively. Furthermore, let $n_+ = |\mathcal{P}|$, $n_- = |\mathcal{N}|$. With this notation in place, $\Delta(\mathbf{z}, \mathbf{y})$ for PRBEP is defined as $b/n_+$ if $b = c$ and $-\infty$ otherwise (Joachims, 2005).

ROCArea, on the other hand, measures how many pairs of examples are mis-ordered. Denote $m = n_+ n_-$. (Joachims, 2005) proposed using the following empirical risk, $R_{\text{emp}}$, to directly optimize the ROCArea:

$$\frac{1}{m}\max_{\mathbf{z}\in\{-1,1\}^m}\left[\sum_{i\in\mathcal{P}, j\in\mathcal{N}} \tfrac{1}{2}(1 - z_{ij}) + z_{ij}\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j)\right]. \tag{3}$$

The empirical risks in (2) and (3) are non-smooth and this leads to difficulties in solving (1). However, cutting plane methods such as SVM-Perf (Joachims, 2006) and BMRM (Teo et al., 2010) can handle such problems. At each iteration these algorithms only require a sub-gradient of $R_{\text{emp}}$, which can be efficiently computed by a *separation* algorithm with $O(n \log n)$ effort for both (2) and (3) (Joachims, 2005). One can show

that cutting plane methods can find an $\epsilon$ accurate solution of (1) after computing $O(\frac{1}{\lambda\epsilon})$ sub-gradients (Teo et al., 2010). These rates are optimal and cannot be improved (Zhang et al., 2011).

One possible approach to break the $\Omega(\frac{1}{\lambda\epsilon})$ barrier is to approximate (1) by a smooth function, which in turn can be efficiently minimized by using either an accelerated gradient method or a quasi-Newton method (Nesterov, 2005, 2007). This technique for non-smooth optimization was pioneered by Nesterov (2005). We now describe some relevant details. The necessary mathematical preliminaries can be found in Appendix B.

### 1.1 Nesterov's Formulation[1]

Let $A$ be a linear transform and assume that we can find a smooth function $g^*_\mu(A^\top \mathbf{w})$ with a Lipschitz continuous gradient such that $\left|R_{\text{emp}}(\mathbf{w}) - g^*_\mu(A^\top \mathbf{w})\right| \leq \mu$ for all $\mathbf{w}$. It is easy to see that

$$J_\mu(\mathbf{w}) := \frac{\lambda}{2}\|\mathbf{w}\|^2 + g^*_\mu(A^\top \mathbf{w}) \tag{4}$$

satisfies $|J_\mu(\mathbf{w}) - J(\mathbf{w})| \leq \mu$ for all $\mathbf{w}$. In particular, if we set $\mu = \epsilon/2$ and find a $\mathbf{w}'$ such that $J_\mu(\mathbf{w}') \leq \min_{\mathbf{w}} J_\mu(\mathbf{w}) + \epsilon/2$, then it follows that $J(\mathbf{w}') \leq \min_{\mathbf{w}} J(\mathbf{w}) + \epsilon$. In other words, $\mathbf{w}'$ is an $\epsilon$ accurate solution for (1).

If we apply Nesterov's accelerated gradient method (Nesterov, 1983) to $J_\mu(\mathbf{w})$, as shown in Appendix A, one can find an $\epsilon$ accurate solution to $J(\mathbf{w})$ by querying the gradient of $g^\star_\mu(A^\top \mathbf{w})$ for

$$O^*\left(\sqrt{D}\,\|A\| \min\left\{\frac{1}{\epsilon}, \frac{1}{\sqrt{\lambda\epsilon}}\right\}\right) \tag{5}$$

number of times (Nesterov, 2005). Here $\|A\|$ is the matrix norm of $A$, and $D$ is a geometric constant that depends solely on $g^*_\mu$ and is independent of $\epsilon$ or $\lambda$.

Compared with the $O(\frac{1}{\lambda\epsilon})$ rates of cutting plane methods, the $\frac{1}{\sqrt{\lambda\epsilon}}$ part in (5) is already superior. Furthermore, many applications require $\lambda \ll \epsilon$ and in this case the $\frac{1}{\epsilon}$ part of the rate is even better. Note cutting plane methods rely on $\frac{\lambda}{2}\|\mathbf{w}\|^2$ to stabilize each update, and so they often converge slowly when $\lambda$ is small (Do et al., 2009).

Although the above scheme is conceptually simple, the smoothing of the objective function in (1) has to be performed very carefully in order to avoid dependence on $n$, the size of the training set. The main difficulties are two-fold. First, one needs to obtain a smooth approximation $g^*_\mu(A^\top \mathbf{w})$ to $R_{\text{emp}}(\mathbf{w})$ such that $\sqrt{D}\,\|A\|$ is small (ideally a constant). Second, we need to show that computing the gradient of $g^*_\mu(A^\top \mathbf{w})$ is no harder than computing a sub-gradient of $R_{\text{emp}}(\mathbf{w})$. In the sequel we will demonstrate how both the above difficulties can be overcome. Before describing our scheme in detail we would like to place our work in context by discussing some relevant related work.

[1] For completeness we reproduce technical details from Nesterov (2005) in Appendix A.

### 1.2 Related Work

Training large models by using variants of stochastic gradient descent has recently become increasingly popular (Bottou, 2008; Shalev-Shwartz et al., 2007). However, stochastic gradient descent can only be applied when the empirical risk is *additively decomposable*, that is, it can be written as the average loss over individual data points. Since the non-linear multivariate scores such as the ones that we consider in this paper are not additively decomposable, this rules out the application of online algorithms to these problems.

Traditionally, batch optimizers such as the popular Sequential Minimal Optimization (SMO) worked in the dual (Platt, 1998). Recently, there has been significant research interest in optimizers which directly optimize (1) because there are some distinct advantages (Teo et al., 2010). Chapelle (2007) observed that to find a $\mathbf{w}$ which generalizes well, one only needs to solve the primal problem to very low accuracy (*e.g.*, $\epsilon \approx 0.01$). In fact, Chapelle (2007) introduced the idea of smoothing the objective function to the machine learning community. Specifically, he proposed to approximate the binary hinge loss by a smooth Huber's loss and used the Newton's method to solve this smoothed problem. This approach yielded the best overall performance in the Wild Competition Track of Sonnenburg et al. (2008) for training binary linear SVMs on large datasets. A similar smoothing approach is proposed by Zhou et al. (2010), but it is also only for binary hinge loss.

However, the smoothing proposed by Chapelle (2007) for the binary hinge loss is rather ad-hoc, and does not easily generalize to (2) and (3). Moreover, a function can be smoothed in many different ways and (Chapelle, 2007) did not explicitly relate the influence of smoothing on the rates of convergence of the solver. In contrast, we propose principled approaches to overcome these problems.

Of course, other smoothing techniques have also been explored in the literature. A popular approach is to replace the nonsmooth max term by a smooth log-sum-exp approximation (Boyd & Vandenberghe, 2004). In the case of binary classification this approximation is closely related to logistic regression (Bartlett et al., 2006; Zhang, 2004), and is equivalent to using an entropy regularizer in the dual. However, as we discuss in Section 2.1.2 this technique has some undesirable properties.

### 1.3 Notation and Paper Outline

We assume a standard setup as in Nesterov (2005), and make a running assumption that all $\mathbf{x}_i$ reside in a Euclidean ball of radius $R$. In Section 2 we will discuss how the smoothing function $g^*_\mu(A^\top \mathbf{w})$ can be designed for (2) and (3). We will focus on efficiently computing the gradient of the smooth objective function in Section 3. Empirical evaluation is presented in Section 4, and the paper concludes with a discussion in Section 5.

## 2 Reformulating the Empirical Risk

In order to approximate $R_{\text{emp}}$ by $g^*_\mu$ we will write $R_{\text{emp}}(\mathbf{w})$ as $g^*(A^\top \mathbf{w})$ for an appropriate linear transform $A$ and convex function $g^*$ with domain $Q$. Let $d$ be a strongly convex function with modulus 1 defined on $Q$. Furthermore, assume $\min_{\boldsymbol{\beta}\in Q} d(\boldsymbol{\beta}) = 0$ and denote $D = \max_{\boldsymbol{\beta}\in Q} d(\boldsymbol{\beta})$. $d$ is called a *prox-function*. Set

$$g^\star_\mu = (g + \mu d)^\star.$$

Then, one can show that $g^\star_\mu(A^\top \mathbf{w})$ is smooth and its gradient is Lipschitz continuous with constant at most $\frac{1}{\mu}\|A\|^2$. Clearly,

$$|g^\star_\mu(A^\top \mathbf{w}) - R_{\text{emp}}(\mathbf{w})| < \mu D, \tag{6}$$

and by choosing $\mu = \epsilon/D$, we can guarantee the approximation is uniformly upper bounded by $\epsilon$.

There are indeed many different ways of writing $R_{\text{emp}}(\mathbf{w})$ as $g^*(A^\top \mathbf{w})$, but the next two sections will demonstrate the advantage of our design.

### 2.1 Contingency Table Based Loss

Letting $\mathcal{S}^k$ denote the $k$ dimensional probability simplex, we can rewrite (2) as:

$$\begin{aligned}
R_{\text{emp}}(\mathbf{w}) &= \max_{\mathbf{z}\in\{-1,1\}^n} \left[\Delta(\mathbf{z},\mathbf{y}) + \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i\rangle (z_i - y_i)\right] \\
&= \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \sum_{\mathbf{z}\in\{-1,1\}^n} \alpha_{\mathbf{z}} \left(\Delta(\mathbf{z},\mathbf{y}) + \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i\rangle (z_i - y_i)\right) \quad (7) \\
&= \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \frac{-2}{n}\sum_{i=1}^n y_i \langle \mathbf{w}, \mathbf{x}_i\rangle \left(\sum_{\mathbf{z}:z_i=-y_i} \alpha_{\mathbf{z}}\right) + \sum_{\mathbf{z}\in\{-1,1\}^n} \alpha_{\mathbf{z}}\Delta(\mathbf{z},\mathbf{y}).
\end{aligned}$$

Define $\beta_i = \sum_{\mathbf{z}:z_i=-y_i} \alpha_{\mathbf{z}}$, then it is not hard to show that $R_{\text{emp}}(\mathbf{w})$ can be further rewritten as

$$\max_{\boldsymbol{\beta}\in[0,1]^n} \left\{\frac{-2}{n}\sum_{i=1}^n y_i \langle \mathbf{w}, \mathbf{x}_i\rangle \beta_i - g(\boldsymbol{\beta})\right\} \text{ where} \tag{8}$$

$$g(\boldsymbol{\beta}) := -\max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{\mathbf{z}} \alpha_{\mathbf{z}}\Delta(\mathbf{z},\mathbf{y}). \tag{9}$$

Here $\mathcal{A}$ is a subset of $\mathcal{S}^{2^n}$ defined via $\mathcal{A} = \left\{\boldsymbol{\alpha} \text{ s.t. } \sum_{\mathbf{z}:z_i=-y_i} \alpha_{\mathbf{z}} = \beta_i \text{ for all } i\right\}$. Indeed, this rewriting only requires that the mapping from $\boldsymbol{\alpha} \in \mathcal{S}^{2^n}$ to $\boldsymbol{\beta} \in Q := [0,1]^n$ is surjective. This is clear because for any $\boldsymbol{\beta} \in [0,1]^n$, a pre-image $\boldsymbol{\alpha}$ can be constructed:

$$\boldsymbol{\alpha}_{\mathbf{z}} = \prod_{i=1}^n \gamma_i, \quad \text{where} \quad \gamma_i = \begin{cases} \beta_i & \text{if } z_i = -y_i \\ 1-\beta_i & \text{if } z_i = y_i. \end{cases}$$

Furthermore we can show $g(\boldsymbol{\beta})$ is convex on $\boldsymbol{\beta} \in [0,1]^n$, (see Appendix C for a proof). Using (8) it immediately follows that $R_{\text{emp}}(\mathbf{w}) = g^\star(A^\top \mathbf{w})$ where $A$ is a $p$-by-$n$ matrix whose $i$-th column is $\frac{-2}{n} y_i \mathbf{x}_i$, and $g^*$ denotes the Fenchel dual of $g$.

#### 2.1.1 $\sqrt{D}\,\|A\|$ for our design

Let us choose the prox-function $d(\boldsymbol{\beta})$ as $\frac{1}{2}\|\boldsymbol{\beta}\|^2$. Then $D = \max_{\boldsymbol{\beta}\in[0,1]^n} d(\boldsymbol{\beta}) = \frac{n}{2}$. The norm of $A = \frac{-2}{n}(y_1\mathbf{x}_1, \ldots, y_n\mathbf{x}_n)$ can be tightly upper bounded by $\frac{2}{n}\sqrt{n}R = \frac{2R}{\sqrt{n}}$. Hence

$$\sqrt{D}\,\|A\| \le \sqrt{\frac{n}{2}}\frac{2R}{\sqrt{n}} = \sqrt{2}R.$$

#### 2.1.2 Alternatives

It is illuminating to see how naive choices for smoothing $R_{\text{emp}}$ can lead to large values of $\sqrt{D}\,\|A\|$. For instance, by (7), $R_{\text{emp}}(\mathbf{w})$ can be written as $h^\star(B^\top \mathbf{w})$ where $h(\boldsymbol{\alpha}) = -n\sum_{\mathbf{z}\in\{-1,1\}^n} \Delta(\mathbf{z},\mathbf{y})\alpha_{\mathbf{z}}$ if $\boldsymbol{\alpha}_{\mathbf{z}} \in [0, n^{-1}]$ and $\sum_{\mathbf{z}} \alpha_{\mathbf{z}} = \frac{1}{n}$, and $\infty$ elsewhere. $B$ is a $p$-by-$2^n$ matrix whose $\mathbf{z}$-th column is $\sum_{i=1}^n \mathbf{x}_i(z_i - y_i)$. $h(\boldsymbol{\alpha})$ has exactly the same form as the matrix game objective in (Nesterov, 2005), and a natural choice of prox-function $d$ is the entropy $d(\boldsymbol{\alpha}) = \sum_{\mathbf{z}} \alpha_{\mathbf{z}} \ln \alpha_{\mathbf{z}} + \frac{1}{n}\log n + \log 2$. However one can show that in this case $\sqrt{D}\,\|A\|$ can be $\Omega(nR)$ which grows linearly with $n$, the number of training examples. Similarly, the smoothing scheme proposed by Zhang et al. (2011) also suffers from a linearly growing $\sqrt{D}\,\|A\|$.

Conceptually the key difficulty arises because the entropy $d$ is defined on a $2^n$ dimensional simplex. However, one can bypass the $\Omega(nR)$ dependence when $\Delta$ is additively decomposable. For example, if $\Delta(\mathbf{z},\mathbf{y}) = \frac{1}{n}\sum_i \delta(z_i \neq y_i)$ in (2), then one can define $d(\boldsymbol{\alpha}) = \sum_i \alpha_i \log \alpha_i + (1-\alpha_i)\log(1-\alpha_i)$. By a straightforward derivation (omitted for brevity), one can show that $g^\star_\mu(A^\top \mathbf{w})$ recovers the logistic loss with its slope controlled by $\mu$, and hence $\sqrt{D}\,\|A\|$ is constant. However, since our $\Delta$ is not decomposable, the log-sum-exp approximation to (2) is not advantageous.

### 2.2 ROCArea

We rewrite $R_{\text{emp}}(\mathbf{w})$ from (3) as:

$$\frac{1}{m}\max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^m}}\sum_{\mathbf{z}\in\{-1,1\}^m}\alpha_{\mathbf{z}}\left[\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{N}}\frac{1}{2}(1-z_{ij})+z_{ij}\mathbf{w}^\top(\mathbf{x}_i-\mathbf{x}_j)\right]$$
$$=\frac{1}{2}+\frac{1}{m}\max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\sum_{i,j}\left(\sum_{\mathbf{z}}z_{ij}\alpha_{\mathbf{z}}\right)\right. \tag{10}$$
$$\left.+\sum_{i,j}\mathbf{w}^\top(\mathbf{x}_i-\mathbf{x}_j)\left(\sum_{\mathbf{z}}z_{ij}\alpha_{\mathbf{z}}\right)\right].$$

Let us define $\beta_{ij}=\sum_{\mathbf{z}}z_{ij}\alpha_{\mathbf{z}}$ for all $(i,j)\in\mathcal{P}\times\mathcal{N}$. This yields a compact form of $R_{\text{emp}}(\mathbf{w})$:

$$\frac{1}{2}+\frac{1}{m}\max_{\boldsymbol{\beta}}\left[-\frac{1}{2}\sum_{i,j}\beta_{ij}+\sum_{i,j}\beta_{ij}\mathbf{w}^\top(\mathbf{x}_i-\mathbf{x}_j)\right]. \tag{11}$$

Clearly $\beta_{ij}\in[-1,1]$. In fact, we can further show that the mapping from $\boldsymbol{\alpha}\in\mathcal{S}^{2^m}$ to $\boldsymbol{\beta}\in Q:=[-1,1]^m$ is surjective. For any $\boldsymbol{\beta}$, a (non-unique) pre-image $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}_{\mathbf{z}}=\prod_{ij}\gamma_{ij},\quad\text{where}\quad\gamma_{ij}=\begin{cases}\frac{1}{2}(1+\beta_{ij}) & \text{if } z_{ij}=1\\ \frac{1}{2}(1-\beta_{ij}) & \text{if } z_{ij}=-1\end{cases}.$$

Ignoring $\frac{1}{2}$, and using (11) $R_{\text{emp}}(\mathbf{w})$ can be written as $g^\star(A^\top\mathbf{w})$ where

$$g(\boldsymbol{\beta})=\begin{cases}\frac{1}{2m}\sum_{i,j}\beta_{ij} & \text{if } \boldsymbol{\beta}\in[-1,1]^m\\ +\infty & \text{elsewhere}\end{cases},$$

and $A$ is a $p$-by-$m$ matrix whose $(ij)$-th column is $\frac{1}{m}(\mathbf{x}_i-\mathbf{x}_j)$ for all $(i,j)\in\mathcal{P}\times\mathcal{N}$.

#### 2.2.1 $\sqrt{D}\,\|A\|$ for our design

Choose prox-function $d(\boldsymbol{\beta})=\frac{1}{2}\sum_{i,j}\beta_{ij}^2$. By a simple calculation, $D=\max_{\boldsymbol{\beta}}d(\boldsymbol{\beta})=\frac{m}{2}$. We can upper bound the norm of $A$ by

$$\begin{aligned}\|A\|&=\max_{\|\mathbf{w}\|=\|\boldsymbol{\beta}\|=1}\mathbf{w}^\top A\boldsymbol{\beta}\\&=\max_{\|\mathbf{w}\|=\|\boldsymbol{\beta}\|=1}\sum_{i,j}\frac{\beta_{ij}}{m}\mathbf{w}^\top(\mathbf{x}_i-\mathbf{x}_j)\\&\le\max_{\|\boldsymbol{\beta}\|=1}\frac{2R}{m}\sum_{i,j}|\beta_{ij}|\le 2Rm^{-\frac{1}{2}},\end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. Therefore

$$\sqrt{D}\,\|A\|\le\sqrt{\frac{m}{2}}\cdot\frac{2R}{\sqrt{m}}=\sqrt{2}R.$$

## 3 Computing the Gradient Efficiently

The last building block required to make the whole scheme work is an efficient algorithm to compute the gradient of the smoothed empirical risk $g_\mu^\star(A^\top\mathbf{w})$. By the chain rule and Corollary X.1.4.4 of (Hiriart-Urruty & Lemaréchal, 1993), we have

$$\frac{\partial}{\partial\mathbf{w}}g_\mu^\star(A^\top\mathbf{w})=A\boldsymbol{\beta}^*\text{ where} \tag{12}$$
$$\boldsymbol{\beta}^*=\operatorname*{argmax}_{\boldsymbol{\beta}\in Q}\left\langle\boldsymbol{\beta},A^\top\mathbf{w}\right\rangle-g(\boldsymbol{\beta})-\mu d(\boldsymbol{\beta}). \tag{13}$$

Two major difficulties arise in computing the above gradient: (13) can be hard to solve (*e.g.* in the case of contingency table based loss), and the matrix vector product in (12) can be costly (*e.g.* $O(n^2p)$ for ROCArea). Below we show how these operations can be performed in $O(n\log n)$ time.

### 3.1 Contingency table based loss

Since $A$ is a $p\times n$ dimensional matrix and $\boldsymbol{\beta}^*$ is a $n$ dimensional vector, the matrix vector product in (12) can be computed in $O(np)$ time. Below we focus on solving (13).

To take into account the constraints in the definition of $g(\boldsymbol{\beta})$, we introduce Lagrangian multipliers $\theta_i$ and the optimization in (13) becomes

$$g_\mu^\star(A^\top\mathbf{w})=\max_{\boldsymbol{\beta}\in[0,1]^n}\left\{\frac{-2}{n}\sum_{i=1}^n y_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle\beta_i-\frac{\mu}{2}\sum_{i=1}^n\beta_i^2\right.$$
$$\left.+\max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}}\left[\sum_{\mathbf{z}}\alpha_{\mathbf{z}}\Delta(\mathbf{z},\mathbf{y})+\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\sum_{i=1}^n\theta_i\left(\sum_{\mathbf{z}:z_i=-y_i}\alpha_{\mathbf{z}}-\beta_i\right)\right]\right\}$$
$$\Leftrightarrow\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left\{\max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}}\sum_{\mathbf{z}}\alpha_{\mathbf{z}}\left[\Delta(\mathbf{z},\mathbf{y})+\sum_i\theta_i\delta(z_i=-y_i)\right]\right.$$
$$\left.+\max_{\boldsymbol{\beta}\in[0,1]^n}\sum_{i=1}^n\left(\frac{-\mu}{2}\beta_i^2-\left(\frac{2}{n}y_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle+\theta_i\right)\beta_i\right)\right\}$$
$$\Leftrightarrow\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left\{\max_{\mathbf{z}}\overbrace{\left[\Delta(\mathbf{z},\mathbf{y})+\sum_i\theta_i\delta(z_i=-y_i)\right]}^{:=q(\mathbf{z},\boldsymbol{\theta})}\right. \tag{14}$$
$$\left.+\sum_{i=1}^n\underbrace{\max_{\beta_i\in[0,1]}\left[\frac{-\mu}{2}\beta_i^2-\left(\frac{2}{n}y_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle+\theta_i\right)\beta_i\right]}_{:=h_i(\theta_i)}\right\}.$$

The last step is because all $\beta_i$ are decoupled and can be optimized independently. Let

$$D(\boldsymbol{\theta}):=\max_{\mathbf{z}}q(\mathbf{z},\boldsymbol{\theta})+\sum_{i=1}^n h_i(\theta_i)\text{ and }\boldsymbol{\theta}^*=\operatorname*{argmin}_{\boldsymbol{\theta}}D(\boldsymbol{\theta}).$$

Given $\boldsymbol{\theta}^*$ and denoting $a_i = \frac{-2}{n} y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$, we can recover the optimal $\beta(\theta_i^*)$ from the definition of $h_i(\theta_i^*)$ as follows:

$$\beta_i^* = \beta_i(\theta_i^*) = \begin{cases} 0 & \text{if } \theta_i^* \geq a_i \\ 1 & \text{if } \theta_i^* \leq a_i - \mu \\ \frac{1}{\mu}(a_i - \theta_i^*) & \text{if } \theta_i^* \in [a_i - \mu, a_i] \end{cases} \tag{15}$$

So, the main challenge that remains is to compute $\boldsymbol{\theta}^*$. Towards this end, first note that[2]:

$$\nabla_{\theta_i} h_i(\theta_i) = -\beta_i(\theta_i) \text{ and}$$
$$\nabla_{\boldsymbol{\theta}} q(\mathbf{z}, \boldsymbol{\theta}) = \operatorname{co}\left\{ \delta_{\mathbf{z}} : \mathbf{z} \in \operatorname*{argmax}_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}) \right\}.$$

Here $\delta_{\mathbf{z}} := (\delta(z_1 = -y_1), \ldots, \delta(z_n = -y_n))^\top$ and $\operatorname{co}(\cdot)$ denotes the convex hull of a set. By the first order optimality conditions $\mathbf{0} \in D(\boldsymbol{\theta}^*)$ which implies that

$$-\boldsymbol{\beta}^* \in \operatorname{co}\left\{ \delta_{\mathbf{z}} : \mathbf{z} \in \operatorname*{argmax}_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}) \right\}.$$

The next theorem characterizes $\boldsymbol{\theta}^*$.

**Property 1.** *There must exist a unique optimal solution $\boldsymbol{\theta}^*$ of (20). Furthermore, $\theta_i^* \in [a_i - \mu, a_i]$ and can be computed in $O(n \log n)$ time for PRBEP.*

The proof of the theorem is technical and relegated to Appendix D. The entire algorithm is described in detail in Appendix E.

### 3.2 ROCArea loss

For the ROCArea loss, given the optimal $\boldsymbol{\beta}^*$ in (13) one can compute

$$\frac{\partial}{\partial \mathbf{w}} g_\mu^\star(A^\top \mathbf{w}) = \frac{1}{m} \sum_{i,j} \beta_{ij}^* (\mathbf{x}_i - \mathbf{x}_j)$$
$$= \frac{1}{m} \left[ \sum_{i \in \mathcal{P}} \mathbf{x}_i \underbrace{\left( \sum_{j \in \mathcal{N}} \beta_{ij}^* \right)}_{:=\gamma_i} - \sum_{j \in \mathcal{N}} \mathbf{x}_j \underbrace{\left( \sum_{i \in \mathcal{P}} \beta_{ij}^* \right)}_{:=\gamma_j} \right].$$

If we can efficiently compute all $\gamma_i$ and $\gamma_j$, then the gradient can be computed in $O(np)$ time.

Given $\beta_{ij}^*$, a brute-force approach to compute $\gamma_i$ and $\gamma_j$ takes $O(m)$ time. We exploit the structure of the problem to reduce this cost to $O(n \log n)$, thus matching the complexity of the separation algorithm in (Joachims, 2005). Towards this end, we specialize (13) to ROCArea and write

$$\max_{\boldsymbol{\beta}} \left( \frac{1}{m} \sum_{i,j} \beta_{ij} \mathbf{w}^\top (\mathbf{x}_i - \mathbf{x}_j) - \frac{1}{2m} \sum_{i,j} \beta_{ij} - \frac{\mu}{2} \sum_{i,j} \beta_{ij}^2 \right).$$

[2]We abuse notation slightly and use $\nabla$ to denote both the gradient and sub-gradient

Since all $\beta_{ij}$ are decoupled, their optimal value can be easily found:

$$\beta_{ij}^* = \operatorname{median}(1, a_i - a_j, -1) \text{ where}$$
$$a_i = \frac{1}{\mu m} \left( \mathbf{w}^\top \mathbf{x}_i - \frac{1}{4} \right), \text{ and } a_j = \frac{1}{\mu m} \left( \mathbf{w}^\top \mathbf{x}_j + \frac{1}{4} \right).$$

Below we give a high level description of how $\gamma_i$ for $i \in \mathcal{P}$ can be computed; the scheme for computing $\gamma_j$ for $j \in \mathcal{N}$ is identical. We omit the details for bevity.

For a given $i$, suppose we can divide $\mathcal{N}$ into three sets $\mathcal{M}_i^+$, $\mathcal{M}_i$, and $\mathcal{M}_i^-$ such that

- $j \in \mathcal{M}_i^+ \implies 1 < a_i - a_j$, hence $\beta_{ij}^* = 1$
- $j \in \mathcal{M}_i \implies a_i - a_j \in [-1, 1]$, hence $\beta_{ij}^* = a_i - a_j$
- $j \in \mathcal{M}_i^- \implies a_i - a_j < -1$, hence $\beta_{ij}^* = -1$.

Then, clearly

$$\gamma_i = \sum_{j \in \mathcal{N}} \beta_{ij}^* = |\mathcal{M}_i^+| - |\mathcal{M}_i^-| + |\mathcal{M}_i| \, a_i - \sum_{j \in \mathcal{M}_i} a_j.$$

In order to identify the sets $\mathcal{M}_i^+$, $\mathcal{M}_i$, and $\mathcal{M}_i^-$, we first sort both $\{a_i : i \in \mathcal{P}\}$ and $\{a_j : j \in \mathcal{N}\}$. We then walk down the sorted lists to identify for each $i$ the first and last indices $j$ such that $a_i - a_j \in [-1, 1]$. This is very similar to the algorithm used to merge two sorted lists, and takes $O(n_- + n_+) = O(n)$ time and space. The rest of the operations for computing $\gamma_i$ can be performed in $O(1)$ time with some straightforward book-keeping. The overall complexity of our algorithm is dominated by the complexity of sorting the two lists, which is $O(n \log n)$.

## 4 Empirical Evaluation

We used 11 publicly available datasets and focused our study on two aspects: The reduction in objective value as a function of CPU time, and the generalization performance of the models obtained via the two schemes.

**Practical Considerations** Optimizing the smooth objective function $J_\mu(\mathbf{w})$ using the optimization scheme described in (Nesterov, 2005) requires estimating the Lipschitz constant of the gradient of the $g_\mu^*(A^\top \mathbf{w})$. Although it can automatically tuned by, *e.g.* (Beck & Teboulle, 2009), extra costs are incurred which slows down the optimization empirically (**?**, Appendix G)ee Appendix H). Therefore, we chose to optimize our smooth objective function using L-BFGS, a widely used quasi-Newton solver (Nocedal & Wright, 2006). The L-BFGS code is obtained from `http://www.chokkan.org/software/liblbfgs/`, which is a C port of the original Fortran implementation of L-BFGS by Nocedal. The size of the L-BFGS buffer determines the number of past parameter and gradient

Table 1: Dataset statistics. $n$: #examples, $d$: #features, $s$: feature density.

| dataset | $n$ | $d$ | $s(\%)$ | dataset | $n$ | $d$ | $s(\%)$ | dataset | $n$ | $d$ | $s(\%)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| adult9 | 32,561 | 123 | 11.28 | covertype | 522,911 | 6,274,932 | 22.22 | web8 | 45,546 | 579,586 | 4.24 |
| astro-ph | 62,369 | 99,757 | 0.077 | news20 | 15,960 | 7,264,867 | 0.033 | worm | 615,620 | 804 | 25 |
| aut-avn | 56,862 | 20,707 | 0.25 | real-sim | 57,763 | 2,969,737 | 0.25 | kdd99 | 4,898,431 | 127 | 12.86 |
| reuters-c11 | 23,149 | 1,757,801 | 0.16 | reuters-ccat | 23,149 | 1,757,801 | 0.16 | | | | |

displacement vectors that are used in the construction of the quasi-Newton direction. We set the buffer size to 6. Following Chapelle (2007) we set $\epsilon = 0.001$ and observed that the solution obtained with this approximation generalizes well in most cases.

We compared this scheme with BMRM[3], a state-of-the-art cutting plane method for optimizing multivariate performance scores which directly minimizes the non-smooth objective function $J(\mathbf{w})$ (Teo et al., 2010). We obtained the latest BMRM code from `http://users.rsise.anu.edu.au/~chteo/BMRM.html` and used default settings. For a fair comparison, our smoothed loss was implemented as a subroutine in BMRM and L-BFGS was added as an alternative solver to the BMRM framework. All our code was written in C++ and will be made publicly available.

**Datasets** Table 1 summarizes the datasets used in our experiments. `adult9`, `astro-ph`, `news20`, `real-sim`, `reuters-c11`, `reuters-ccat` are from the same source as in Hsieh et al. (2008). `aut-avn` classifies documents on auto and aviation and was obtained from `http://www.cs.umass.edu/~mccallum/data/sraa.tar.gz`. `covertype` is from UCI repository (Merz & Murphy, 1998). We divided the whole dataset into training, validation and test set in the same way as in (Teo et al., 2010). For all datasets we used the $\lambda$ which yielded the best generalization performance using their corresponding validation sets.

Due to lack of space, we will only present results for three representative datasets in the main body of the paper. Complete results can be found in Appendix G.

### 4.1 Results

**Optimizing the primal objective $J(\mathbf{w})$** In the first experiment we study the effect of $\mu$ on optimizing the primal objective $J(\mathbf{w})$. The choice of $\mu$ is dictated by two conflicting requirements. On the one hand the uniform deviation bound (6) suggests setting $\mu = \epsilon/D$. However, this estimate is very conservative because in (6) we use $D$ an upper bound on the prox-function. In practice, the quality of the approximation depends on the value of the prox-function around the optimum. On the other hand, as $\mu$ increases, the strong convexity of $g$ increases, and this makes $g^*_\mu$ and hence $J_\mu$ easier to optimize. We set $\hat{\mu} = \epsilon/D$ and let $\mu \in \{\hat{\mu}, 100\hat{\mu}, 1000\hat{\mu}\}$ and compare the performance of our scheme with BMRM in Figures 1 and 2.

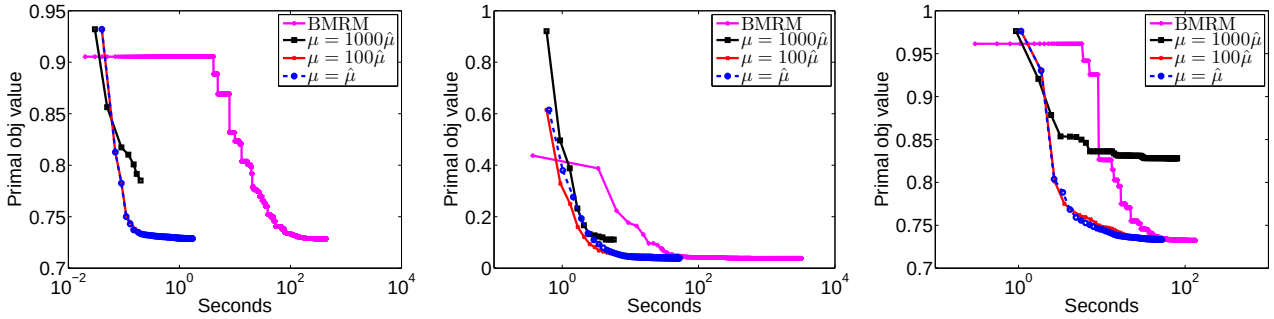
It is clear that for all the values of $\mu$, optimizing the smoothed objective function converges significantly faster than BMRM. Furthermore, for $\mu = \hat{\mu}$ and $\mu = 100\hat{\mu}$ we obtained a solution which was at most $\epsilon$ distance away from the solution obtained by BMRM. Somewhat surprisingly, the optimization trajectories were near identical for $\mu = \hat{\mu}$ and $\mu = 100\hat{\mu}$ indicating that increasing the strong convexity of $g$ did not significantly impact the convergence rates. However, $\mu = 1000\hat{\mu}$ did converge significantly faster, but to a worse quality solution.

**Performance on Test Set** We also studied the evolution of the PRBEP and ROCArea performance on the test data. For this, we obtained the solution after each iteration, computed its performance on the test set, and plotted the results in Figures 3 and 4. Clearly, the intermediate models output by our scheme achieve comparable (or better) PRBEM scores and ROCArea in time orders of magnitude faster those generated by BMRM.

## 5 Conclusion and Discussion

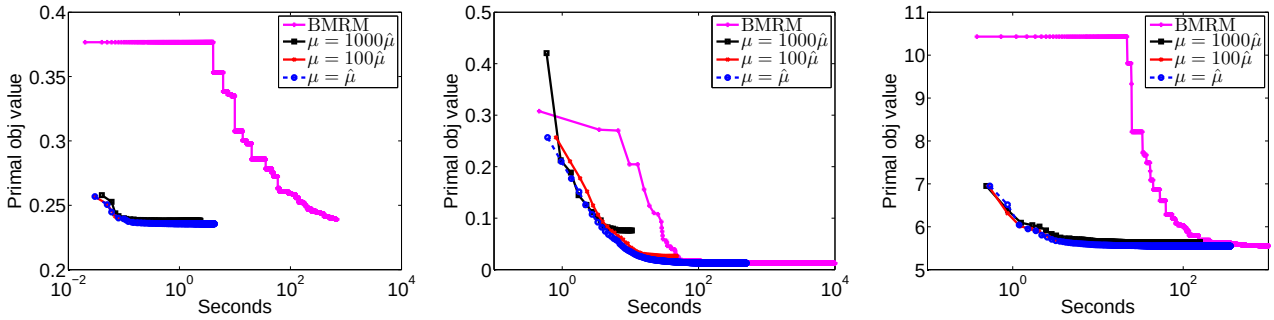
The non-smoothness of the loss function is an important consideration for algorithms which employ the kernel trick (Schölkopf & Smola, 2002). This is because such algorithms typically operate in the dual, and the non-smooth losses lead to sparse dual solutions. In many applications such as natural language processing, the kernel trick is not needed because the input data is sufficiently high dimensional. However, now we are "stuck" with a non-smooth objective function in the primal. While a lot of past work has been devoted to solving this non-smooth problem, one must bear in mind that optimization is a means to an end in machine learning. In line with this philosophy, we proposed efficient smoothing techniques to approximate the non-smooth function. When combined with a smooth optimization algorithm, our technique outperforms state-of-the-art non-smooth optimization algorithms for multivariate performance scores not only in terms of CPU time but also in terms of generaliza-

[3]For quadratic regularizers, BMRM and SVM-Perf are equivalent.




(a) adult9 (b) news20 (c) covertype

Figure 1: Primal objective versus CPU time for PRBEP.




(a) adult9 (b) news20 (c) covertype

Figure 2: Primal objective versus CPU time for ROCArea.

tion performance.

It is also worthwhile noting that smoothing is not the right approach for every non-smooth problem. For example, although it is easy to smooth the $L_1$ norm regularizer, it is not recommended; the sparsity of the solution is an important statistical property of these algorithms and smoothing destroys this property.

In future work we would like to extend our techniques to handle more complicated contingency based multivariate performance measures such as the $F_1$-score. We would also like to extend smoothing to matching loss functions commonly used in ranking, where we believe our techniques will solve a smoothed version of the Hungarian marriage problem.

## References


Bartlett, Peter L., Jordan, Michael I., and McAuliffe, Jon D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473): 138–156, 2006.

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bottou, Léon. Stochastic gradient SVMs. http://leon.bottou.org/projects/sgd, 2008.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

Chapelle, Olivier. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

Do, C., Le, Q., and Foo, C.S. Proximal regularization for online and batch learning. In *International Conference on Machine Learning*, 2009.

Hiriart-Urruty, J. B. and Lemaréchal, C. *Convex Analysis and Minimization Algorithms, I and II*, volume 305 and 306. Springer-Verlag, 1993.

Hsieh, Cho Jui, Chang, Kai Wei, Lin, Chih Jen, Keerthi, S. Sathiya, and Sundararajan, S. A dual coordinate descent method for large-scale linear SVM. In Cohen, William, McCallum, Andrew, and Roweis, Sam (eds.), *ICML*, pp. 408–415. ACM, 2008.

Joachims, T. A support vector method for multivariate performance measures. In *Proc. Intl. Conf. Machine Learning*, pp. 377–384, 2005.

Joachims, T. Training linear SVMs in linear time. In *Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 217–226, 2006.

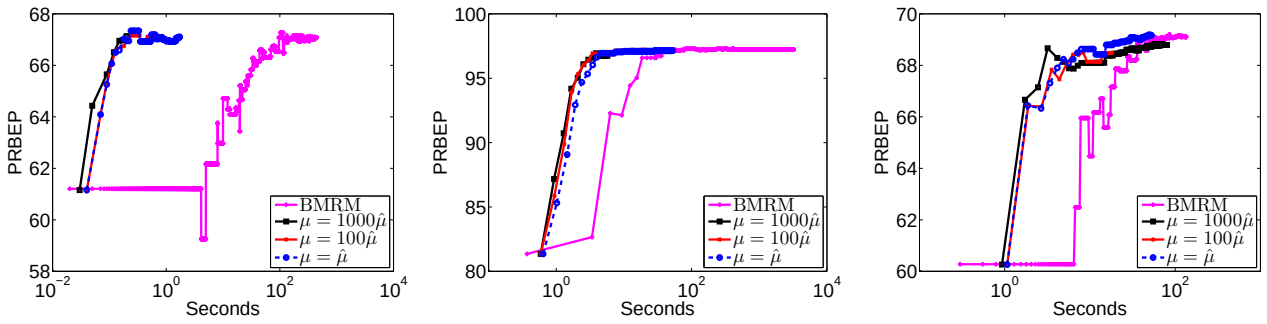
Merz, C. J. and Murphy, P. M. UCI repository of machine learning databases, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Soviet Math. Docl.*, 269:543–547, 1983.

Nesterov, Y. Gradient methods for minimizing composite objective function. Technical Report 76, CORE Discussion Paper, UCL, 2007.

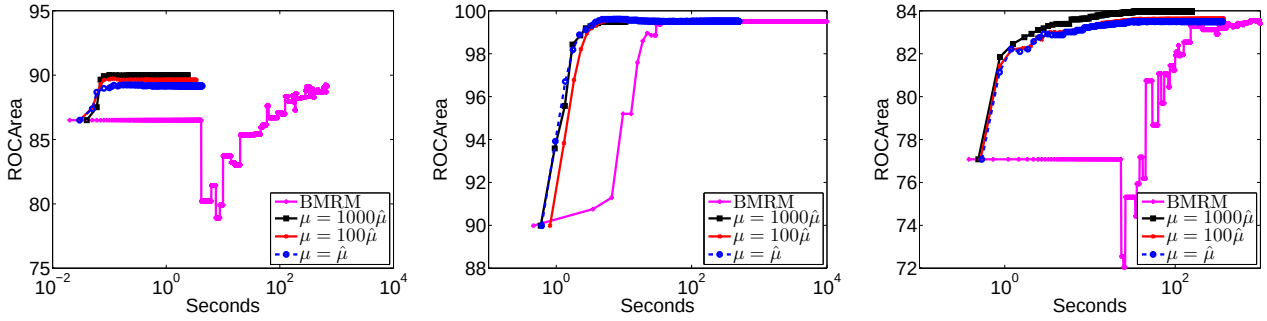
Nesterov, Yurii. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

Nocedal, Jorge and Wright, Stephen J. *Numerical Optimization*. Springer Series in Operations Research. Springer, 2nd edition, 2006.




(a) adult9 (b) news20 (c) covertype

Figure 3: PRBEP on test data versus CPU time.




(a) adult9 (b) news20 (c) covertype

Figure 4: ROCArea on test data versus CPU time.

Platt, J. C. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.

Schölkopf, B. and Smola, A. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proc. Intl. Conf. Machine Learning*, pp. 807–814, 2007.

Sonnenburg, Soeren, Franc, Vojtech, Yom-Tov, Elad, and Sebag, Michele. Pascal large scale learning challenge. 2008. URL http://largescale.ml.tu-berlin.de/workshop/.

Teo, C. H., Vishwanthan, S. V. N., Smola, A. J., and Le, Q. V. Bundle methods for regularized risk minimization. *J. Mach. Learn. Res.*, 11:311–365, January 2010.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

Zhang, Xinhua, Saha, Ankan, and Vishwanathan, S. V. N. Lower bounds on rate of convergence of cutting plane methods. In *Advances in Neural Information Processing Systems 23*, 2011.

Zhou, Tianyi, Tao, Dacheng, and Wu, Xindong. NESVM: a fast gradient method for support vector machines. In *Proc. Intl. Conf. Data Mining*, 2010.

## A The Smoothing Procedure

The idea of the smoothing technique in (Nesterov, 2005) can be motivated by using the Theorem 4.2.1 and 4.2.2 in (Hiriart-Urruty & Lemaréchal, 1993).

**Lemma 1.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and $\nabla f$ is Lipschitz continuous with constant $L$ (called $L$-l.c.g), then $f^\star$ is strongly convex with modulus $\frac{1}{L}$ (called $\frac{1}{L}$-sc). Conversely, if $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is $\sigma$-sc, then $f^\star$ is finite on $\mathbb{R}^n$ and is $\frac{1}{\sigma}$-l.c.g.*

Since $g + \mu d$ is $\mu$-sc, Lemma 1 implies $g^\star_\mu$ is $\frac{1}{\mu}$-l.c.g. By chain rule, one can show that $g^\star_\mu(A^\top \mathbf{w})$ is $L_\mu$-l.c.g where $L_\mu \leq \frac{1}{\mu}\|A\|^2$. Further, the definition of Fenchel dual implies the following uniform deviation bound:

$$g^\star(\mathbf{u}) - \mu D \leq g^\star_\mu(\mathbf{u}) \leq g^\star(\mathbf{u}), \quad \forall\ \mathbf{u} \in \mathbb{R}^n. \tag{16}$$

Hence to find an $\epsilon$ accurate solution to $J(\mathbf{w})$, it suffices to set the maximum deviation $\mu D < \frac{\epsilon}{2}$ (*i.e.* $\mu < \frac{\epsilon}{2D}$), and then find a $\frac{\epsilon}{2}$ accurate solution to $J_\mu$ in (4). Initialize $\mathbf{w}$ to $\mathbf{0}$ and apply Nesterov's accelerated gradient method in (Nesterov, 2007) to $J_\mu$, this takes at most

$$k = \min\left\{\sqrt{\frac{4L_\mu\Delta_0}{\epsilon}}, \ln\frac{L_\mu\Delta_0}{\epsilon} \Big/ \ln\left(1 - \sqrt{\lambda/L_\mu}\right)\right\}$$

number of steps where $\Delta_0 = \frac{1}{2}\|\mathbf{w}^*\|^2$ and $\mathbf{w}^*$ is the minimizer of $J(\mathbf{w})$. Each step involves one gradient query of $g^\star_\mu(A^\top \mathbf{w})$ and some cheap updates. Plugging

in $L_\mu \leq \frac{2D}{\epsilon}\|A\|^2$ and using $\ln(1+\delta) \approx \delta$ when $\delta \approx 0$, we get the iteration bound in (5).

## B Preliminaries of Convex Analysis

The following three notions will be used extensively:

**Definition 1.** *A convex function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strongly convex (s.c.) wrt norm $\|\cdot\|$ if there exists a constant $\sigma > 0$ such that $F - \frac{\sigma}{2}\|\cdot\|^2$ is convex. $\sigma$ is called the modulus of strong convexity of $F$, and for brevity we will call $F$ $\sigma$-strongly convex or $\sigma$-s.c..*

**Definition 2.** *A function $F$ is said to have Lipschitz continuous gradient (l.c.g) if there exists a constant $L$ such that*

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\| \quad \forall\ \mathbf{w}, \mathbf{w}'. \tag{17}$$

*For brevity, we will call $F$ $L$-l.c.g..*

**Definition 3.** *The Fenchel dual of a function $F : E_1 \to E_2$, is a function $F^\star : E_2^\star \to E_1^\star$ given by*

$$F^\star(\mathbf{w}^\star) = \sup_{\mathbf{w} \in E_1} \{\langle \mathbf{w}, \mathbf{w}^\star \rangle - F(\mathbf{w})\} \tag{18}$$

## C Convexity of $g(\beta)$

**Lemma 2.** *$g(\boldsymbol{\beta})$ is convex on $\boldsymbol{\beta} \in [0,1]^n$.*

*Proof.* For any $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ in $[0,1]^n$, suppose their argmax in (9) is attained at (among others) $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ respectively. So $\sum_{\mathbf{z}:z_i=-y_i} \alpha_{\mathbf{z}}^{(j)} = \beta_i^{(j)}$ for $j = 1, 2$. For any $\lambda \in [0,1]$, consider $\boldsymbol{\beta} := \lambda\boldsymbol{\beta}^{(1)} + (1-\lambda)\boldsymbol{\beta}^{(2)}$. Define $\boldsymbol{\alpha} := \lambda\boldsymbol{\alpha}^{(1)} + (1-\lambda)\boldsymbol{\alpha}^{(2)}$. Then clearly $\boldsymbol{\alpha}$ satisfies

$$\sum_{\mathbf{z}:z_i=-y_i} \alpha_{\mathbf{z}} = \lambda\beta_i^{(1)} + (1-\lambda)\beta_i^{(2)} = \beta_i.$$

Therefore $\boldsymbol{\alpha}$ is admissible in the definition of $g(\boldsymbol{\beta})$ and

$$\begin{aligned} g(\boldsymbol{\beta}) &\leq -\sum_{\mathbf{z}} \alpha_{\mathbf{z}} \Delta(\mathbf{z}, \mathbf{y}) \\ &= -\lambda \sum_{\mathbf{z}} \alpha_{\mathbf{z}}^{(1)} \Delta(\mathbf{z}, \mathbf{y}) - (1-\lambda) \sum_{\mathbf{z}} \alpha_{\mathbf{z}}^{(2)} \Delta(\mathbf{z}, \mathbf{y}) \\ &= \lambda g(\boldsymbol{\beta}^{(1)}) + (1-\lambda) g(\boldsymbol{\beta}^{(2)}). \end{aligned}$$

■

## D Proof of Property 1

Let $\hat{\boldsymbol{\theta}}$ be an arbitrary optimal solution and we consider two cases.

- $\hat{\theta}_i > a_i$: In this case reducing $\theta_i$ to $a_i$ does not change $h_i(\theta_i)$. Furthermore, since for all $\mathbf{z}$, $q(\mathbf{z}, \boldsymbol{\theta})$ is monotonically non-decreasing in $\theta_i$, therefore none of the pieces will be increased by reducing $\theta_i$. In summary, replacing $\theta_i$ by $a_i$ also gives an optimal solution.

- $\hat{\theta}_i < a_i - \mu$: In this case increasing $\theta_i$ to $a_i - \mu$ will reduce $h_i(\theta_i)$ by $a_i - \mu - \hat{\theta}_i$. As for the piecewise linear part, since for each $\mathbf{z}$ this change of $\theta_i$ will increase $q(\mathbf{z}, \boldsymbol{\theta})$ by at most $a_i - \mu - \hat{\theta}_i$, hence the whole piecewise linear part is increased by at most $a_i - \mu - \hat{\theta}_i$. So overall, the objective value $D(\boldsymbol{\theta})$ will not increase.

Note that in conjunction with the strong convexity of $h_i(\theta_i)$ in $[a_i - \mu, a_i]$, the optimal solution $\hat{\boldsymbol{\theta}}$ is unique.

# E Details of the algorithm for PRBEP measure

In this section, we change the notations slightly from the main body of the paper for the convenience of presenting technical details. Assume $y_i^*$ is the true label corresponding to $\mathbf{x}_i$.

We write the empirical loss as follows:

$$R_{\text{emp}}(\mathbf{w}) = \max_{\mathbf{y}\in\{-1,1\}^n} \left[ \Delta(\mathbf{y}, \mathbf{y}^*) + \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i \rangle (y_i - y_i^*) \right]$$
$$= \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \sum_{\mathbf{y}\in\{-1,1\}^n} \alpha_{\mathbf{y}} \left( \Delta(\mathbf{y}, \mathbf{y}^*) + \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i \rangle (y_i - y_i^*) \right)$$
$$= \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \frac{-2}{n} \sum_{i=1}^n y_i^* \langle \mathbf{w}, \mathbf{x}_i \rangle \left( \sum_{\mathbf{y}: y_i = -y_i^*} \alpha_{\mathbf{y}} \right) + \sum_{\mathbf{y}\in\{-1,1\}^n} \alpha_{\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}^*).$$

Define $\beta_i = \sum_{\mathbf{y}: y_i = -y_i^*} \alpha_{\mathbf{y}}$, then it is not hard to show that $R_{\text{emp}}(\mathbf{w})$ can be further rewritten as

$$\max_{\boldsymbol{\beta}\in[0,1]^n} \left\{ \frac{-2}{n} \sum_{i=1}^n y_i^* \langle \mathbf{w}, \mathbf{x}_i \rangle \beta_i - g(\boldsymbol{\beta}) \right\} \quad \text{where}$$
$$g(\boldsymbol{\beta}) := - \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}: \forall i, \sum_{\mathbf{y}: y_i = -y_i^*} \alpha_{\mathbf{y}} = \beta_i} \sum_{\mathbf{y}} \alpha_{\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}^*).$$

Indeed, this rewriting only requires that the mapping from $\boldsymbol{\alpha} \in \mathcal{S}^{2^n}$ to $\boldsymbol{\beta} \in [0,1]^n$ is surjective. This is clear because for any $\boldsymbol{\beta} \in [0,1]^n$, a pre-image $\boldsymbol{\alpha}$ can be constructed as

$$\boldsymbol{\alpha}_{\mathbf{y}} = \prod_{i=1}^n \gamma_i, \quad \text{where} \quad \gamma_i = \begin{cases} \beta_i & \text{if } y_i = -y_i^* \\ 1 - \beta_i & \text{if } y_i = y_i^* \end{cases}.$$

Furthermore we can show $g$ is convex (and obviously closed).

**Lemma 3.** *$g(\boldsymbol{\beta})$ is convex on $\boldsymbol{\beta} \in [0,1]^n$.*

*Proof.* For any $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ in $[0,1]^n$, suppose their argmax in the definition of $g$ is attained at (among others) $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ respectively. So $\sum_{\mathbf{y}: y_i = -y_i^*} \alpha_{\mathbf{y}}^{(j)} = \beta_i^{(j)}$ for $j = 1, 2$. For any $\lambda \in [0,1]$, consider $\boldsymbol{\beta} := \lambda\boldsymbol{\beta}^{(1)} + (1-\lambda)\boldsymbol{\beta}^{(2)}$. Define $\boldsymbol{\alpha} := \lambda\boldsymbol{\alpha}^{(1)} + (1-\lambda)\boldsymbol{\alpha}^{(2)}$. Then clearly $\boldsymbol{\alpha}$ satisfies

$$\sum_{\mathbf{y}: y_i = -y_i^*} \alpha_{\mathbf{y}} = \lambda\beta_i^{(1)} + (1-\lambda)\beta_i^{(2)} = \beta_i.$$

Therefore $\boldsymbol{\alpha}$ is admissible in the definition of $g(\boldsymbol{\beta})$ and

$$g(\boldsymbol{\beta}) \le -\sum_{\mathbf{y}} \alpha_{\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}^*)$$
$$= -\lambda \sum_{\mathbf{y}} \alpha_{\mathbf{y}}^{(1)} \Delta(\mathbf{y}, \mathbf{y}^*) - (1-\lambda) \sum_{\mathbf{y}} \alpha_{\mathbf{y}}^{(2)} \Delta(\mathbf{y}, \mathbf{y}^*)$$
$$= \lambda g(\boldsymbol{\beta}^{(1)}) + (1-\lambda) g(\boldsymbol{\beta}^{(2)}). \qquad \blacksquare$$

So $R_{\text{emp}}(\mathbf{w})$ can be written as $g^\star(A^\top \mathbf{w})$ where $A$ is a $d$-by-$n$ matrix whose $i$-th column is $\frac{-2}{n} y_i^* \mathbf{x}_i$.

**Rate of convergence.** Let us use prox-function $d(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n \beta_i^2$. Then $D = \max_{\boldsymbol{\beta}} d(\boldsymbol{\beta}) = \frac{n}{2}$. Endowing $L_2$ norm on both the $\mathbf{w}$ and $\boldsymbol{\beta}$ space, the norm of $A$ can be upper bounded by $\frac{2}{n}\sqrt{n}R = \frac{2R}{\sqrt{n}}$. Hence

$$D \|A\|^2 \le \frac{n}{2} \frac{4R^2}{n} = 2R^2.$$

Plugging this into Nesterov's convergence expression, we obtain that to find an $\epsilon$ accurate solution, one needs

$$O^*\left( R \min\left\{ \frac{1}{\epsilon}, \frac{1}{\sqrt{\lambda\epsilon}} \right\} \right)$$

number of iterations. This is independent of $n$. Note if we apply entropy regularizer directly on $\boldsymbol{\alpha}$, then the above iteration complexity will be multiplied by $n$.

**Computing the gradient.** Again, $\frac{\partial}{\mathbf{w}} g_\mu^\star(A^\top \mathbf{w}) = \frac{-2}{n}\sum_{i=1}^n \hat{\beta}_i y_i^* \mathbf{x}_i$, where $\hat{\beta}_i$ is the optimal solution of

$$g_\mu^\star(A^\top \mathbf{w}) = \max_{\boldsymbol{\beta}\in[0,1]^n} \left\langle \boldsymbol{\beta}, A^\top \mathbf{w} \right\rangle - g(\boldsymbol{\beta}) - \frac{\mu}{2}\sum_{i=1}^n \beta_i^2. \tag{19}$$

To take into account the constraints in the definition of $g(\boldsymbol{\beta})$, we introduce Lagrangian multipliers $\theta_i$ and

the optimization in (19) is equivalent to

$$
\begin{aligned}
g_\mu^\star(A^\top \mathbf{w}) = & \max_{\boldsymbol{\beta}\in[0,1]^n} \left\{ \frac{-2}{n}\sum_{i=1}^n y_i^* \langle \mathbf{w}, \mathbf{x}_i\rangle \beta_i - \frac{\mu}{2}\sum_{i=1}^n \beta_i^2 \right. \\
& \left. + \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \left[ \sum_{\mathbf{y}} \alpha_{\mathbf{y}} \Delta(\mathbf{y},\mathbf{y}^*) + \min_{\boldsymbol{\theta}\in\mathbb{R}^n} \sum_{i=1}^n \theta_i \left( \sum_{\mathbf{y}:y_i=-y_i^*} \alpha_{\mathbf{y}} - \beta_i \right) \right] \right\} \\
\Leftrightarrow \min_{\boldsymbol{\theta}\in\mathbb{R}^n} & \left\{ \max_{\boldsymbol{\alpha}\in\mathcal{S}^{2^n}} \sum_{\mathbf{y}} \alpha_{\mathbf{y}} \left[ \Delta(\mathbf{y},\mathbf{y}^*) + \sum_i \theta_i \delta(y_i = -y_i^*) \right] \right. \\
& \left. + \max_{\boldsymbol{\beta}\in[0,1]^n} \sum_{i=1}^n \left( \frac{-\mu}{2}\beta_i^2 - \left( \frac{2}{n} y_i^* \langle \mathbf{w}, \mathbf{x}_i\rangle + \theta_i \right) \beta_i \right) \right\} \\
\Leftrightarrow \min_{\boldsymbol{\theta}\in\mathbb{R}^n} & \left\{ \max_{\mathbf{y}} \overbrace{\left[ \Delta(\mathbf{y},\mathbf{y}^*) + \sum_i \theta_i \delta(y_i = -y_i^*) \right]}^{:=q(\mathbf{y},\boldsymbol{\theta})} \right. \\
& \left. + \sum_{i=1}^n \underbrace{\max_{\beta_i\in[0,1]} \left[ \frac{-\mu}{2}\beta_i^2 - \left( \frac{2}{n} y_i^* \langle \mathbf{w}, \mathbf{x}_i\rangle + \theta_i \right) \beta_i \right]}_{:=h_i(\theta_i)} \right\}.
\end{aligned} \tag{20}
$$

The last step is because all $\beta_i$ are decoupled and can be optimized independently. Denoting $a_i = \frac{-2}{n} y_i^* \langle \mathbf{w}, \mathbf{x}_i\rangle$, we have

$$
h_i(\theta_i) = \begin{cases} 0 & \text{if } \theta_i \geq a_i \\ -\theta_i + a_i - \frac{\mu}{2} & \text{if } \theta_i \leq a_i - \mu \\ \frac{1}{2\mu}(\theta_i - a_i)^2 & \text{if } \theta_i \in [a_i - \mu, a_i] \end{cases} \tag{21}
$$

with the optimal $\beta_i$ being

$$
\beta_i(\theta_i) = \begin{cases} 0 & \text{if } \theta_i \geq a_i \\ 1 & \text{if } \theta_i \leq a_i - \mu \\ \frac{1}{\mu}(a_i - \theta_i) & \text{if } \theta_i \in [a_i - \mu, a_i] \end{cases}. \tag{22}
$$

So we only need to find the optimal $\hat{\boldsymbol{\theta}}$ to (20):

$$
\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}} D(\boldsymbol{\theta}), \text{ where } D(\boldsymbol{\theta}) := \max_{\mathbf{y}} q(\mathbf{y},\theta) + \sum_{i=1}^n h_i(\theta_i).
$$

Then compute $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\theta}})$ according to (22), and get the gradient. Our major tool for finding $\hat{\boldsymbol{\theta}}$ is the first order condition: $\mathbf{0}$ must be a subgradient of $D(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$. So we compute the gradient of $h_i$ ($h_i$ is differentiable):

$$
\nabla h_i(\theta_i) = \begin{cases} 0 & \text{if } \theta_i \geq a_i \\ -1 & \text{if } \theta_i \leq a_i - \mu \\ \frac{1}{\mu}(\theta_i - a_i) & \text{if } \theta_i \in [a_i - \mu, a_i] \end{cases} = -\beta_i(\theta_i).
$$

Therefore, $\beta_i(\theta_i)$ can be viewed as the *height* of gradient, *i.e.* how much is the gradient below 0. The subdifferential of the piecewise linear part can be computed using the following lemma.

**Lemma 4.** *Let $f(\boldsymbol{\theta}) = \max_{i\in\mathcal{I}} \{\boldsymbol{\theta}^\top \mathbf{a}_i + b_i\}$ where $\mathcal{I}$ is an index set. Let $\mathcal{J}$ be the index set of* argmax*: $\mathcal{J} = \{j \in \mathcal{I} : \boldsymbol{\theta}^\top \mathbf{a}_i + b_i = f(\boldsymbol{\theta})\}$. Then the subdifferential*

$$
\partial f(\boldsymbol{\theta}) = \left\{ \sum_{j\in\mathcal{J}} \alpha_j \mathbf{a}_j : \boldsymbol{\alpha} \in \mathcal{S}^{|\mathcal{J}|} \right\}. \tag{23}
$$

So the optimization of (20) is boiled down to finding a $\boldsymbol{\theta}$ such that there is an element in (23) which cancels with the gradient of $h_i$. The major challenge is that the $\operatorname{argmax}_{\mathbf{y}}$ in (20) can have a lot of tie.

The following property gives an important characterization of the solution to (20).

**Property 2.** *There must exist an optimal solution to (20) which lies in $\prod_{i=1}^n [a_i - \mu, a_i]$. In conjunction with the strong convexity of $h_i(\theta_i)$ in $[a_i - \mu, a_i]$, there must exist a unique optimal solution $\hat{\boldsymbol{\theta}}$ in $\prod_{i=1}^n [a_i - \mu, a_i]$.*

*Proof.* Let $\hat{\boldsymbol{\theta}}$ be an arbitrary optimal solution and we consider two cases. i) if $\hat{\theta}_i > a_i$, then reducing $\theta_i$ to $a_i$ does not change $h_i(\theta_i)$. Furthermore, since for all $\mathbf{y}$, $q(\mathbf{y},\boldsymbol{\theta})$ is monotonically non-decreasing in $\theta_i$, therefore none of the pieces will be increased by reducing $\theta_i$. In summary, replacing $\theta_i$ by $a_i$ also gives an optimal solution. ii) if $\hat{\theta}_i < a_i - \mu$, then increasing $\theta_i$ to $a_i - \mu$ will reduce $h_i(\theta_i)$ by $a_i - \mu - \hat{\theta}_i$. As for the piecewise linear part, since for each $\mathbf{y}$ this change of $\theta_i$ will increase $q(\mathbf{y},\boldsymbol{\theta})$ by at most $a_i - \mu - \hat{\theta}_i$, hence the whole piecewise linear part is increased by at most $a_i - \mu - \hat{\theta}_i$. So overall, the objective $D$ value will not increase. ■

In the sequel, we will use $\hat{\boldsymbol{\theta}}$ to denote the unique optimal solution in $\prod_{i=1}^n [a_i - \mu, a_i]$, and denote $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\theta}})$. By Lemma 4 and the form of $q(\mathbf{y},\theta)$, we have

**Lemma 5.** *$\hat{\beta}_i = 0$ if, and only if, all $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ satisfy $\hat{y}_i = y_i^*$. $\hat{\beta}_i = 1$ if, and only if, all $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ satisfy $\hat{y}_i = -y_i^*$. $\hat{\beta}_i \in (0,1)$ if, and only if, there exists $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ such that $\hat{y}_i = -y_i^*$, and there exists $\hat{\mathbf{y}}' \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ such that $\hat{y}_i' = y_i^*$.*

*Proof.* Since $\mathbf{0} \in \partial D(\hat{\boldsymbol{\theta}})$, so there must exist a subgradient $\mathbf{g}$ of $\max_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ such that $g_i = -\nabla h_i(\hat{\theta}_i) = \hat{\beta}_i = 0$. Hence by Lemma 4, for all $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$ we have $\frac{\partial}{\partial\theta_i} q(\mathbf{y},\boldsymbol{\theta}) = 0$. But note that for any fixed $\mathbf{y}$,

$$
\frac{\partial}{\partial\theta_i} q(\mathbf{y},\boldsymbol{\theta}) = \delta(y_i = -y_i^*).
$$

Therefore $\hat{y}_i = y_i^*$ for all $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y},\hat{\boldsymbol{\theta}})$. The other results in this lemma can be proved similarly. ■

**Characterizing the solution for contingency table based loss.** As in the previous subsection, without loss of generality suppose there are $n_+$ positive examples which are associated with $\mathbf{x}_i^+$, $y_i^+$, $a_i^+$, $\hat{\theta}_i^+$, $\hat{\beta}_i^+$, and $h_i^+(\cdot)$, and $n_-$ negative examples which are associated with $\mathbf{x}_j^-$, $y_j^-$, $a_j^-$, $\hat{\theta}_j^-$, $\hat{\beta}_j^-$, and $h_j^-(\cdot)$. So $n = n_+ + n_-$. We will always use $i$ to index positive examples and $j$ to index negative examples. Also assume both $\{a_i^+\}$ and $\{a_j^-\}$ are already sorted *decreasingly*:

$$a_1^+ \geq a_2^+ \ldots \geq a_{n_+}^+, \quad \text{and} \quad a_1^- \geq \ldots \geq a_{n_-}^-.$$

We can further characterize the optimal solution $\hat{\boldsymbol{\theta}}$ by making use of the fact that $\Delta(\mathbf{y}, \mathbf{y}^*)$ is based on the contingency table, *i.e.* $\Delta(\mathbf{y}, \mathbf{y}^*) = \Delta(b(\mathbf{y}), c(\mathbf{y}))$ where

$$b(\mathbf{y}) = \sum_{i=1}^{n_+} \delta(y_i^+ = -1), \text{ and } c(\mathbf{y}) = \sum_{j=1}^{n_-} \delta(y_j^- = 1). \quad (24)$$

The first property is that the optimal $\left\{\hat{\theta}_i^+\right\}$ and $\left\{\hat{\theta}_j^-\right\}$ are both in a decreasing order.

**Lemma 6.** *$\hat{\theta}_i^+ \geq \hat{\theta}_{i'}^+$ for all $i, i' \in [n_+]$ and $i < i'$. $\hat{\theta}_j^- \geq \hat{\theta}_{j'}^-$ for all $j, j' \in [n_-]$ and $j < j'$.*

*Proof.* We just prove the first clause since the second can be proved similarly. If $a_i^+ = a_{i'}^+$, then by symmetry $\hat{\theta}_i^+ = \hat{\theta}_{i'}^+$. So let us assume $a_i^+ > a_{i'}^+$. Suppose the lemma were not true, *i.e.* there exist $i, i' \in [n_+]$ and $i < i'$ such that $\hat{\theta}_i^+ < \hat{\theta}_{i'}^+$. Let us swap the value of $\hat{\theta}_i^+$ and $\hat{\theta}_{i'}^+$, and call the new $\boldsymbol{\theta}$ as $\bar{\boldsymbol{\theta}}$. Below we show that $D(\bar{\boldsymbol{\theta}}) < D(\hat{\boldsymbol{\theta}})$ which contradicts with the optimality of $\hat{\boldsymbol{\theta}}$. Clearly, $\max_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \max_{\mathbf{y}} q(\mathbf{y}, \bar{\boldsymbol{\theta}})$. However, it is not hard to see that by $a_i^+ > a_{i'}^+$ and $\hat{\theta}_i^+ < \hat{\theta}_{i'}^+$,

$$\begin{aligned} &- a_i^+ \hat{\theta}_i^+ - a_{i'}^+ \hat{\theta}_{i'}^+ - (-a_i^+ \hat{\theta}_{i'}^+ - a_{i'}^+ \hat{\theta}_i^+) \\ =&(a_i^+ - a_{i'}^+)(\hat{\theta}_{i'}^+ - \hat{\theta}_i^+) > 0. \end{aligned}$$

So

$$\begin{aligned} &h_i^+(\hat{\theta}_i^+) + h_{i'}^+(\hat{\theta}_{i'}^+) = \frac{1}{2\mu}\left[(a_i^+ - \hat{\theta}_i^+)^2 + (a_{i'}^+ - \hat{\theta}_{i'}^+)^2\right] \\ &> \frac{1}{2\mu}\left[(a_i^+ - \hat{\theta}_{i'}^+)^2 + (a_{i'}^+ - \hat{\theta}_i^+)^2\right] = h_i^+(\bar{\theta}_i^+) + h_{i'}^+(\bar{\theta}_{i'}^+). \end{aligned}$$

$\hat{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ match on all other $h_i^+$ and $h_j^-$. So $D(\bar{\boldsymbol{\theta}}) < D(\hat{\boldsymbol{\theta}})$. Contradiction. ■

The second most important property of the optimal solution $\hat{\boldsymbol{\theta}}$ is that its height is monotonically decreasing among the set of positive and negative examples respectively.

**Lemma 7.** *$\hat{\beta}_i^+ \geq \hat{\beta}_{i'}^+$ for all $i, i' \in [n_+]$ and $i < i'$. $\hat{\beta}_j^- \geq \hat{\beta}_{j'}^-$ for all $j, j' \in [n_-]$ and $j < j'$.*

*Proof.* We just prove the first clause since the second can be proved similarly. Suppose otherwise there exist $i, i' \in [n_+]$ and $i < i'$ such that $\hat{\beta}_i^+ < \hat{\beta}_{i'}^+$. We will show $\mathbf{0}$ cannot be a subgradient of $D$ for such a solution. We first claim that for any $\hat{\mathbf{y}}$ which maximizes $q(\mathbf{y}, \hat{\boldsymbol{\theta}})$, $\hat{y}_{i'} = -1$ must entail $\hat{y}_i = -1$. Suppose otherwise $\hat{y}_i = 1$. Then let us swap their values, *i.e.* consider a new assignment $\bar{\mathbf{y}}$ where $\bar{y}_{i'} = 1$ and $\bar{y}_i = -1$. Then $\Delta(\hat{\mathbf{y}}, \mathbf{y}^*) = \Delta(\bar{\mathbf{y}}, \mathbf{y}^*)$ since $\hat{\mathbf{y}}$ and $\bar{\mathbf{y}}$ have the same false positive and false negative. However,

$$\begin{aligned} &q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\bar{\mathbf{y}}, \hat{\boldsymbol{\theta}}) \\ &= \sum_{k=1}^{n_+} \hat{\theta}_k^+ \delta(\hat{y}_k = -1) - \sum_{k=1}^{n_+} \hat{\theta}_k^+ \delta(\bar{y}_k = -1) \\ &= \hat{\theta}_{i'}^+ - \hat{\theta}_i^+ = a_{i'}^+ - \mu\hat{\beta}_{i'}^+ - (a_i^+ - \mu\hat{\beta}_i^+) \\ &= (a_{i'}^+ - a_i^+) + \mu(\hat{\beta}_i^+ - \hat{\beta}_{i'}^+) < 0, \end{aligned}$$

which contradicts with the assumption that $\hat{\mathbf{y}}$ maximizes $q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. Therefore by Lemma 4, for any subgradient $\mathbf{g}$ of $\max_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ we have $g_{i'}^+ \leq g_i^+$. But then $g_{i'}^+ - \hat{\beta}_{i'}^+$ and $g_i^+ - \hat{\beta}_i^+$ cannot equal 0 simultaneously because if so then

$$\hat{\beta}_{i'}^+ = g_{i'}^+ \leq g_i^+ = \hat{\beta}_i^+,$$

which contradicts with our assumption. ■

**Algorithm for solving (20) with PRBEP loss.** Now we specialize the loss $\Delta$ to PRBEP and give a concrete algorithm to find $\hat{\boldsymbol{\theta}}$. For PRBEP loss, $\Delta(\mathbf{y}, \mathbf{y}^*) = -\infty$ if the false negative $b(\mathbf{y})$ in (24) is not equal to the false positive $c(\mathbf{y})$. And when $b(\mathbf{y}) = c(\mathbf{y})$, $\Delta(\mathbf{y}, \mathbf{y}^*)$ is defined as $\frac{1}{n} b(\mathbf{y}) = \frac{1}{n} c(\mathbf{y})$. So we just abbreviate $\Delta$ as $\Delta(k) = k/n$. Our algorithm also works when $\Delta(k)$ satisfies the diminishing gain property: $\Delta(k') - \Delta(k'-1) \leq \Delta(k) - \Delta(k-1)$ for all $k' > k > 0$. But for simplicity, we will stick to $\Delta(k) = k/n$ in the sequel.

According to Lemma 7, the first few $\hat{\beta}_1^+, \ldots \hat{\beta}_k^+$ are all 1, followed by some $\hat{\beta}_{k+1}^+, \ldots \hat{\beta}_{k'}^+$ lying in $(0, 1)$, and finally all the rest $\hat{\beta}_{k'+1}^+, \ldots \hat{\beta}_{n_+}^+$ are straight 0. Of particular importance is the phase transition point $k$ which takes into account both the positive and negative examples. Let $\hat{\theta}_0^+ = \hat{\theta}_0^- = 0$ and define

$$k := \max\left\{0 \leq i \leq \min\{n_+, n_-\} : \hat{\theta}_i^+ + \hat{\theta}_i^- + \frac{1}{n} > 0\right\}. \quad (25)$$

Depending on the value of $k$ and whether $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} = 0$, our discussion can be divided into several cases, which are summarized in Figure 5.

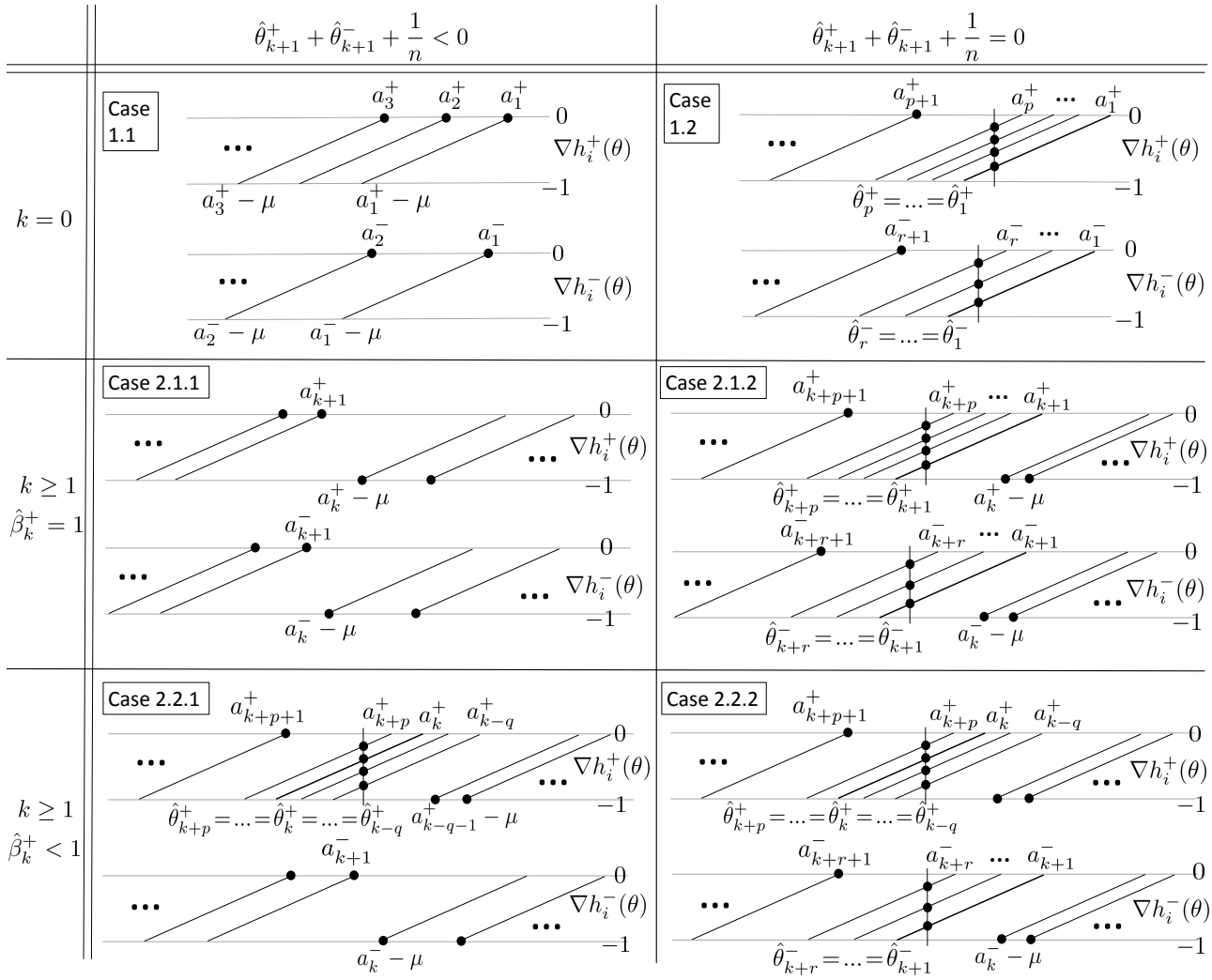



Figure 5: Summary of all cases of the solution of PRBEP loss. The black dot represents the solution $\hat{\theta}^+_i$ and $\hat{\theta}^-_j$.

**Case 1.** Suppose $k = 0$. Then $\hat{\theta}^+_1 + \hat{\theta}^-_1 + \frac{1}{n} \leq 0$.

**Case 1.1** Suppose $\hat{\theta}^+_1 + \hat{\theta}^-_1 + \frac{1}{n} < 0$. By the definition of $\Delta(k)$ and the decreasing order of $\hat{\theta}^+_i$ and $\hat{\theta}^-_j$ in Lemma 6, $\operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ must be the correct labeling $\mathbf{y}^*$, *i.e.* $b(\mathbf{y}) = c(\mathbf{y}) = 0$ and $q(\mathbf{y}^*, \hat{\boldsymbol{\theta}}) = 0$. Therefore $\partial \max_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \{\mathbf{0}\}$. Since $\mathbf{0} \in \partial D(\hat{\boldsymbol{\theta}})$, so for all $i \in [n_+]$, $\hat{\beta}^+_i = -\nabla h^+_i(\hat{\theta}^+_i) = 0$, *i.e.* $\hat{\theta}^+_i = a^+_i$; and for all $j \in [n_-]$, $\hat{\beta}^-_j = -\nabla h^-_j(\hat{\theta}^-_j) = 0$, *i.e.* $\hat{\theta}^-_j = a^-_j$. In summary, the condition of falling in this case is

$$a^+_1 + a^-_1 + \frac{1}{n} \leq 0. \tag{26}$$

Note we allow equality here because it lies on the boundary of cases and does give a proper solution.

**Case 1.2** Suppose $\hat{\theta}^+_1 + \hat{\theta}^-_1 + \frac{1}{n} = 0$. Assume $\hat{\theta}^+_1 = \ldots = \hat{\theta}^+_p > \hat{\theta}^+_{p+1}$ $(p \geq 1)$, and $\hat{\theta}^-_1 = \ldots = \hat{\theta}^-_r > \hat{\theta}^-_{r+1}$ $(r \geq 1)$. Then we have two properties, one for these tied elements of $\hat{\boldsymbol{\theta}}$ (Property 4) and one for the rest elements of $\hat{\boldsymbol{\theta}}$ (Property 3). We first state the latter since it is simpler.

**Property 3.** *For all $i > p$, $\hat{\beta}^+_i = 0$,* i.e. *$\hat{\theta}^+_i = a^+_i$. For all $j > r$, $\hat{\beta}^-_j = 0$,* i.e. *$\hat{\theta}^-_j = a^-_j$.*

*Proof.* We just prove the first clause since the second can be proved similarly. Suppose otherwise there exists $i > p$ and $\hat{\beta}^+_i > 0$. Then by Lemma 5, there must exist a $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_i = -1$. Since $\hat{\theta}_i < \hat{\theta}_p$, so $b(\hat{\mathbf{y}}) \geq p + 1$ and $\hat{y}^+_1 = \ldots = \hat{y}^+_p = -1$. Since $c(\hat{\mathbf{y}}) = b(\hat{\mathbf{y}}) \geq p + 1$, there must exist $j \in [n_-]$ such that $\hat{y}^-_j = 1$. Now construct a new assignment $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}^+_i = 1$ and $\bar{y}^-_j = -1$. Now $\bar{\mathbf{y}}$ commits one less false positive and

false negative than $\hat{\mathbf{y}}$, and by $\hat{\theta}_i < \hat{\theta}_1$ and $\hat{\theta}_j^- \le \hat{\theta}_1^-$,

$$q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\bar{\mathbf{y}}, \hat{\boldsymbol{\theta}}) = \hat{\theta}_i^+ + \hat{\theta}_j^- + \frac{1}{n} < \hat{\theta}_1^+ + \hat{\theta}_1^- + \frac{1}{n} = 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

**Property 4.** $\sum_{i=1}^p \hat{\beta}_i^+ = \sum_{j=1}^r \hat{\beta}_j^-$.

*Proof.* By Property 3 and Lemma 5, all $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ must satisfy $\hat{y}_{p+1}^+ = \ldots = \hat{y}_{n_+}^+ = 1$ and $\hat{y}_{r+1}^- = \ldots = \hat{y}_{n_-}^- = -1$. Noting $\hat{\theta}_1^+ + \hat{\theta}_1^- + \frac{1}{n} = 0$, we can explicitly express the set of $\operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ as

$$\mathcal{Y} = \Big\{\mathbf{y} : \sum_{i=1}^p \delta(y_i^+ = -1) = \sum_{j=1}^r \delta(y_j^- = -1), \\ y_i^+ = 1 \ \forall i > p, \ y_j^- = -1 \ \forall j > r\Big\}.$$

Since $\mathbf{0} \in \partial D(\hat{\boldsymbol{\theta}})$, so there exists a distribution $\boldsymbol{\alpha}$ over $\mathbf{y} \in \mathcal{Y}$ such that for all $i \in [p]$ and $j \in [r]$:

$$\hat{\beta}_i^+ = \sum_{\mathbf{y} \in \mathcal{Y}: y_i^+ = -1} \alpha_{\mathbf{y}}, \quad \text{and} \quad \hat{\beta}_j^- = \sum_{\mathbf{y} \in \mathcal{Y}: y_j^- = 1} \alpha_{\mathbf{y}}.$$

So

$$\sum_{i=1}^p \hat{\beta}_i^+ = \sum_{i=1}^p \sum_{\mathbf{y} \in \mathcal{Y}: y_i^+ = -1} \alpha_{\mathbf{y}} = \sum_{\mathbf{y} \in \mathcal{Y}} \alpha_{\mathbf{y}} \sum_{i=1}^p \delta(y_i^+ = -1)$$

$$= \sum_{\mathbf{y} \in \mathcal{Y}} \alpha_{\mathbf{y}} \sum_{j=1}^r \delta(y_j^- = -1) = \sum_{j=1}^r \sum_{\mathbf{y} \in \mathcal{Y}: y_j^- = 1} \alpha_{\mathbf{y}} = \sum_{j=1}^r \hat{\beta}_j^-.$$

■

Incidentally, the $\boldsymbol{\alpha}$ in the proof of Property 4 can be explicitly constructed as

$$\alpha_{\mathbf{y}} = \prod_{i=1}^p \gamma_i^+ \prod_{j=1}^r \gamma_j^- \text{ where } \gamma_i^+ = \begin{cases} \hat{\beta}_i^+ & \text{if } y_i^+ = -1 \\ 1 - \hat{\beta}_i^+ & \text{if } y_i^+ = 1 \end{cases},$$

$$\gamma_j^- = \begin{cases} \hat{\beta}_j^- & \text{if } y_j^- = 1 \\ 1 - \hat{\beta}_j^- & \text{if } y_j^- = -1 \end{cases}.$$

Property 4 allows us to fix the value of $\hat{\theta}_1^+$ and $\hat{\theta}_1^-$, and to conveniently check whether the optimal $\hat{\boldsymbol{\theta}}$ falls in this case. Using $\hat{\theta}_1^+ + \hat{\theta}_1^- + \frac{1}{n} = 0$, it is obvious that $\hat{\theta}_1^+$ must be the root of the following function

$$G(\theta) = \sum_{i=1}^{n_+} \beta_i^+(\theta) - \sum_{j=1}^{n_-} \beta_j^-\left(-\frac{1}{n} - \theta\right),$$
$$\text{where} \quad \theta \in [a_1^+ - \mu, a_1^+],$$
$$\text{and} \quad -\frac{1}{n} - \theta \in [a_1^- - \mu, a_1^-].$$

$G(\theta)$ is monotonically decreasing. So the necessary and sufficient conditions for falling into this case are: i) the above domain of $\theta$ is not empty; ii) $G(\theta)$ at the maximum of the domain is non-positive and $G(\theta)$ at the minimum of the domain is non-negative.

Once the root $\theta^*$ is found, simply set $p$ to the greatest $i$ such that $a_i^+ \ge \theta^*$, and set $r$ to the greatest $j$ such that $a_j^- \ge -\frac{1}{n} - \theta^*$. Assign $\hat{\theta}_1^+ = \ldots = \hat{\theta}_p^+ = \theta^*$ and $\hat{\theta}_1^- = \ldots = \hat{\theta}_r^- = -\frac{1}{n} - \theta^*$. Then it is easy to see that $\sum_{i=1}^p \beta_i^+(\theta^*) = \sum_{j=1}^r \beta_j^-(-\frac{1}{n} - \theta^*)$.

Algorithmically, the root finding algorithm based on binary search takes $O(\log n)$ steps, but at each step, for a given $\theta$ it may require $O(n)$ time to compute $G^+(\theta) := \sum_{i=1}^{n_+} \beta_i^+(\theta)$ (and $\sum_{j=1}^{n_-} \beta_j^-(\theta)$). This cost can be reduced to $O(\log n)$ (amortized) if we conduct the following $O(n \log n)$ pre-computation. Spend $O(n \log n)$ time sorting $\{a_i^+, a_i^+ - \mu : i \in [n_+]\}$ and then $G^+(\theta)$ is linear between two adjacent points in the sorted list. Next spend $O(n)$ time recording their slopes and the value of $G^+$ at the end points: $\{G^+(a_i^+ - \mu), G^+(a_i^+) : i \in [n_+]\}$. Now given a $\theta$, use binary search (which costs $O(\log n)$) to find the interval of the sorted list which $\theta$ belongs to, and then $G^+(\theta)$ can be computed by using its slope and the value of $G^+$ at the end points.

**Case 2.** Suppose $1 \le k < \min\{n_+, n_-\}$. We start with the following property which will significantly simplify our analysis.

**Property 5.** *$\hat{\beta}_k^+ = 1$ or $\hat{\beta}_k^- = 1$, or both.*

*Proof.* First it is clearly impossible that $\hat{\theta}_k^+ = \hat{\theta}_{k+1}^+$ and $\hat{\theta}_k^- = \hat{\theta}_{k+1}^-$ simultaneously, because then $k+1$ also satisfies the condition in (25). Without loss of generality, assume $\hat{\theta}_k^+ > \hat{\theta}_{k+1}^+$ and below we show $\hat{\beta}_k^+ = 1$. Suppose otherwise $\hat{\beta}_k^+ < 1$. Then by Lemma 5, there must exist a $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_k^+ = 1$. This rules out the possibility that $b(\hat{\mathbf{y}}) \ge k$ because if so then $\hat{y}_k^+ = -1$ as guaranteed by $\hat{\theta}_{k+1}^+ < \hat{\theta}_k^+$. Now that $b(\hat{\mathbf{y}}) = c(\hat{\mathbf{y}}) < k$, there must exist $j \in [k]$ such that $\hat{y}_j^- = -1$. Then we can construct a new $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_k^+ = -1$ and $\bar{y}_j^- = 1$. So $\bar{\mathbf{y}}$ makes one more false positive and false negative, and

$$q(\bar{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} + \hat{\theta}_k^+ + \hat{\theta}_j^- \ge \frac{1}{n} + \hat{\theta}_k^+ + \hat{\theta}_k^- > 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

By Property 5, let us assume $\hat{\beta}_k^- = 1$ and so by Lemma 7 we have for all $j \in [k]$, $\hat{\beta}_j^- = 1$ and $\hat{\theta}_j^- = a_j^- - \mu$. Now let us consider two cases.

**Case 2.1.** Suppose $\hat{\beta}_k^+ = 1$. By Lemma 7 we have $\hat{\beta}_i^+ = 1$ and $\hat{\theta}_i^+ = a_i^+ - \mu$ for all $i \in [k]$. By the definition of $k$, we have $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} \leq 0$. So we further consider two cases.

**Case 2.1.1.** $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} < 0$. We have the following property in this case.

**Property 6.** *For all $i > k$, $\hat{\beta}_i^+ = 0$,* i.e. *$\hat{\theta}_i^+ = a_i^+$. For all $j > k$, $\hat{\beta}_j^- = 0$,* i.e. *$\hat{\theta}_j^- = a_j^-$.*

*Proof.* Let us just prove the first clause since the second one can be proved similarly. Suppose otherwise there exists $i \geq k+1$ and $\hat{\beta}_i^+ > 0$. Then by Lemma 5 there must exist a $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_i = -1$. Since $\hat{\beta}_1^+ = \ldots = \hat{\beta}_k^+ = 1$, Lemma 5 guarantees that $\hat{y}_1^+ = \ldots = \hat{y}_1^+ = -1$. So $c(\hat{\mathbf{y}}) = b(\hat{\mathbf{y}}) \geq k+1$. As a result, there must exist an index $j \geq k+1$ such that $\hat{y}_j^- = 1$. Now construct a new $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_i^+ = 1$ and $\bar{y}_j^- = -1$. So $\bar{\mathbf{y}}$ commits one less false positive and false negative than $\hat{\mathbf{y}}$, and so

$$q(\hat{\mathbf{y}}) - q(\bar{\mathbf{y}}) = \frac{1}{n} + a_i^+ + a_j^- \leq \frac{1}{n} + a_{k+1}^+ + a_{k+1}^- < 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

In summary, the solution in this case is $\hat{\theta}_i^+ = a_i^+ - \mu$, $\forall\ i \leq k$; $\hat{\theta}_i^+ = a_i^+$, $\forall\ i > k$; $\hat{\theta}_j^- = a_j^- - \mu$, $\forall\ j \leq k$; and $\hat{\theta}_j^- = a_j^-$, $\forall\ j > k$. The conditions for falling in this case are also easy to describe and check:

$$\begin{cases} a_{k+1}^+ \leq a_k^+ - \mu, \quad a_{k+1}^- \leq a_k^- - \mu, \\ a_{k+1}^+ + a_{k+1}^- + \frac{1}{n} \leq 0, \\ (a_k^+ - \mu) + (a_k^- - \mu) + \frac{1}{n} \geq 0. \end{cases}$$

Note we allow equality in the last two inequalities because they lie on the boundary of different cases and do give a proper solution.

**Case 2.1.2.** $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} = 0$. Suppose $\hat{\theta}_{k+1}^+ = \hat{\theta}_{k+2}^+ = \ldots = \hat{\theta}_{k+p}^+ > \hat{\theta}_{k+p+1}^+$ $(p \geq 1)$ and $\hat{\theta}_{k+1}^- = \hat{\theta}_{k+2}^- = \ldots = \hat{\theta}_{k+r}^- > \hat{\theta}_{k+r+1}^-$ $(r \geq 1)$. Then it is straightforward to set the value for the "tails".

**Property 7.** *For all $i > k+p$, $\hat{\beta}_i^+ = 0$,* i.e. *$\hat{\theta}_i^+ = a_i^+$. For all $j > k+r$, $\hat{\beta}_j^- = 0$,* i.e. *$\hat{\theta}_j^- = a_j^-$.*

*Proof.* We just prove the first clause since the second can be proved in the same way. Suppose otherwise there exists a $i > k+p$ such that $\hat{\beta}_i^+ > 0$. So by Lemma 5 there must exist $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_i^+ = -1$. Since $\hat{\theta}_i^+ < \hat{\theta}_{k+1}^+$, so $\hat{y}_{i-1}^+ = \ldots = \hat{y}_1^+ = -1$, *i.e.* $c(\hat{\mathbf{y}}) = b(\hat{\mathbf{y}}) \geq i \geq k+2$. Therefore there must exist $j \geq k+2$ such that $\hat{y}_j^- = 1$. Now consider a new assignment $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_i^+ = 1$ and $\bar{y}_j^- = -1$. So $\bar{\mathbf{y}}$ commits one less false positive and false negative, and by $\hat{\theta}_i^+ < \hat{\theta}_{k+1}^+$ and $\hat{\theta}_j^- \leq \hat{\theta}_{k+1}^-$,

$$q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\bar{\mathbf{y}}, \bar{\boldsymbol{\theta}}) = \frac{1}{n} + \hat{\theta}_i^+ + \hat{\theta}_j^- < \frac{1}{n} + \hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- = 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

So the only values to be determined are $\hat{\theta}_{k+1}^+ (= \ldots = \hat{\theta}_{k+p}^+)$ and $\hat{\theta}_{k+1}^- (= \ldots = \hat{\theta}_{k+r}^-)$. To this end, we use a property similar to Property 4.

**Property 8.** $\sum_{i=k+1}^{k+p} \hat{\beta}_i^+ = \sum_{j=k+1}^{k+r} \hat{\beta}_j^-$.

The proof is exactly the same as for Property 4 with the only difference that the set of $\operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ is now

$$\left\{ \mathbf{y} : \sum_{i=k+1}^{k+p} \delta(y_i^+ = -1) = \sum_{j=k+1}^{k+r} \delta(y_j^- = 1), \right.$$
$$y_i^+ = -1\ \forall i \leq k, \text{ and } y_i^+ = 1\ \forall i > k+p,$$
$$\left. y_j^- = 1\ \forall j \leq k, \text{ and } y_j^- = -1\ \forall j > k+r \right\}.$$

Using Property 8, we can not only fix the value of $\hat{\theta}_{k+1}^+$ and $\hat{\theta}_{k+1}^-$, but also check whether the real optimal $\hat{\boldsymbol{\theta}}$ falls in this case. Using $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} = 0$, it is obvious that $\hat{\theta}_{k+1}^+$ must be the root of the following function

$$G_k(\theta) = \sum_{i=k+1}^{n_+} \beta_i^+(\theta) - \sum_{j=k+1}^{n_-} \beta_j^- \left( -\frac{1}{n} - \theta \right),$$
$$\text{where } \theta \in [a_{k+1}^+ - \mu, \min\{a_{k+1}^+, a_k^+ - \mu\}],$$
$$\text{and } -\frac{1}{n} - \theta \in [a_{k+1}^- - \mu, \min\{a_{k+1}^-, a_k^- - \mu\}].$$

$G_k$ is monotonically decreasing and has at most $6n$ linear pieces. To fall in this case, the following conditions are necessary and sufficient. First the above domain of $\theta$ is not empty. Second $G_k(\theta)$ at the maximum of the domain is non-positive and $G_k(\theta)$ at the minimum of the domain is non-negative.

When $G_k(\theta^*) = 0$, set $k+p$ to the greatest $i$ such that $a_i^+ \geq \theta^*$, and set $k+r$ to the greatest $j$ such that $a_j^- \geq -\frac{1}{n} - \theta^*$. Assign $\hat{\theta}_{k+1}^+ = \ldots = \hat{\theta}_{k+p}^+ = \theta^*$ and $\hat{\theta}_{k+1}^- = \ldots = \hat{\theta}_{k+r}^- = -\frac{1}{n} - \theta^*$. Then it is easy to see that $\sum_{i=k+1}^{k+p} \beta_i^+(\theta^*) = \sum_{j=k+1}^{k+r} \beta_j^-(-\frac{1}{n} - \theta^*)$.

The binary search for the root takes $O(\log n)$ steps and each step costs $O(\log n)$ using the same pre-

computation and root finding procedure as for $G(\theta)$ in case 1.2.

**Case 2.2.** Suppose $\hat{\beta}_k^+ < 1$. Using the proof of Property 5, $\hat{\theta}_k^+$ must be tied with $\hat{\theta}_{k+1}^+$ and assume $\hat{\theta}_k^+ = \ldots = \hat{\theta}_{k+p}^+ > \hat{\theta}_{k+p+1}^+$ ($p \geq 1$). Now by the definition of $k$, we have $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} \leq 0$. So we furthermore consider two cases.

**Case 2.2.1.** $\hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- + \frac{1}{n} < 0$. We have the following property in this case.

**Property 9.** *For all $i > k+p$, $\hat{\beta}_i^+ = 0$, i.e. $\hat{\theta}_i^+ = a_i^+$. For all $j \geq k+1$, $\hat{\beta}_j^- = 0$, i.e. $\hat{\theta}_j^- = a_j^-$.*

*Proof.* Note since $\hat{\beta}_k^+ < 1$ and $\hat{\beta}_k^- = 1$, the situations for negative and positive side are not symmetric. To prove the first clause, suppose there exists $i > k+p$ such that $\hat{\beta}_i^+ > 0$. Then by Lemma 5 there must exist a $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_i^+ = -1$. Since $\hat{\theta}_i^+ < \hat{\theta}_{k+p}^+$, so $c(\hat{\mathbf{y}}) = b(\hat{\mathbf{y}}) \geq k+p+1$. Therefore there must exist a $j \geq k+1$ such that $\hat{y}_j^- = 1$. Now consider a new assignment $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_i^+ = 1$ and $\bar{y}_j^- = -1$. So $\bar{\mathbf{y}}$ commits one less false positive and false negative, and so

$$q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\bar{\mathbf{y}}, \bar{\boldsymbol{\theta}}) = \frac{1}{n} + \hat{\theta}_i^+ + \hat{\theta}_j^- \leq \frac{1}{n} + \hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- < 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$.

To prove the second clause, suppose there exist $j \geq k+1$ such that $\hat{\beta}_j^- > 0$. Then by Lemma 5 there must exist a $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_j^- = -1$. Since $\hat{\beta}_1^- = \ldots = \hat{\beta}_k^- = 1$, so Lemma 5 implies $\hat{y}_1^- = \ldots = \hat{y}_k^- = 1$. Therefore $b(\hat{\mathbf{y}}) = c(\hat{\mathbf{y}}) \geq k+1$. So there must exist $i \geq k+1$ such that $\hat{y}_i^+ = -1$. Now consider a new assignment $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_i^+ = 1$ and $\bar{y}_j^- = -1$. So $\bar{\mathbf{y}}$ commits one less false positive and false negative, and so

$$q(\hat{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\bar{\mathbf{y}}, \bar{\boldsymbol{\theta}}) = \frac{1}{n} + \hat{\theta}_i^+ + \hat{\theta}_j^- \leq \frac{1}{n} + \hat{\theta}_{k+1}^+ + \hat{\theta}_{k+1}^- < 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

Now the only values to be determined are $\hat{\theta}_1^+, \ldots, \hat{\theta}_{k+p}^+$. Let us assume that $\hat{\theta}_{k+p}^+ = \ldots = \hat{\theta}_k^+ = \ldots = \hat{\theta}_{k-q}^+ < \hat{\theta}_{k-q-1}^+$ ($q \geq 0$). Then one can easily fix $\hat{\theta}_i^+$ for all $i \in [k-q-1]$.

**Property 10.** *For all $i \in [k-q-1]$, $\hat{\beta}_i^+ = 1$, i.e. $\hat{\theta}_i^+ = a_i^+ - \mu$.*

*Proof.* By Lemma 7, it suffices to show $\hat{\beta}_{k-q-1}^+ = 1$. Suppose otherwise $\hat{\beta}_{k-q-1}^+ < 1$, then there must exist $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ such that $\hat{y}_{k-q-1}^+ = 1$. Since $\hat{\theta}_{k-q}^+ < \hat{\theta}_{k-q-1}^+$, so $c(\hat{\mathbf{y}}) = b(\hat{\mathbf{y}}) \leq k-q-2 \leq k-2$. Therefore there must exist $j \in [k-1]$ such that $\hat{y}_j^- = -1$. Now consider a new assignment $\bar{\mathbf{y}}$ which is the same as $\hat{\mathbf{y}}$ except that $\bar{y}_i^+ = -1$ and $\bar{y}_j^- = 1$. So $\bar{\mathbf{y}}$ commits one more false positive and false negative, and so

$$q(\bar{\mathbf{y}}, \hat{\boldsymbol{\theta}}) - q(\hat{\mathbf{y}}, \bar{\boldsymbol{\theta}}) = \frac{1}{n} + \hat{\theta}_{k-q-1}^+ + \hat{\theta}_j^- \geq \frac{1}{n} + \hat{\theta}_k^+ + \hat{\theta}_k^- > 0,$$

which contradicts with $\hat{\mathbf{y}} \in \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. ■

Finally we use the following property to fix the value of $\hat{\theta}_k^+ (= \hat{\theta}_i^+, \ \forall\ k-q \leq i \leq k+p)$.

**Property 11.** $\sum_{i=k-q}^{k+p} \hat{\beta}_i^+ = q+1$.

*Proof.* By Property 9 and $\hat{\beta}_j^- = 1$ for all $j \in [k]$, we have that any $\hat{\mathbf{y}} \in \mathcal{Y} := \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol{\theta}})$ must satisfy that $\hat{y}_j^- = 1$ for all $j \in [k]$ and $\hat{y}_j^- = -1$ for all $j > k$. So $b(\mathbf{y}) = c(\mathbf{y}) = k$. By Property 10, $\hat{y}_i^+ = -1$ for all $i \leq k-q-1$. By Property 9, $\hat{y}_i^+ = 1$ for all $i > k+p$. Therefore among $\hat{y}_{k-q}^+, \ldots, \hat{y}_{k+p}^+$, there are exactly $q+1$ $-1$'s. And since $\hat{\theta}_{k-q}^+ = \ldots = \hat{\theta}_{k+p}^+$, so this is also a sufficient condition for $\hat{\mathbf{y}}$ to maximize $q(\mathbf{y}, \hat{\boldsymbol{\theta}})$. Hence

$$\sum_{i=k-q}^{k+p} \delta(y_i^+ = -1) = q+1, \quad \forall\ \mathbf{y} \in \mathcal{Y}. \tag{27}$$

Since $\mathbf{0} \in \partial D(\hat{\boldsymbol{\theta}})$, so there exists a distribution $\boldsymbol{\alpha}$ over $\mathbf{y} \in \mathcal{Y}$ such that

$$\hat{\beta}_i^+ = \sum_{\mathbf{y} \in \mathcal{Y}: y_i^+ = -1} \alpha_{\mathbf{y}}, \quad \forall\ i \in \{k-q, \ldots, k+p\}.$$

Then

$$\sum_{i=k-q}^{k+p} \hat{\beta}_i^+ = \sum_{i=k-q}^{k+p} \sum_{\mathbf{y} \in \mathcal{Y}: y_i^+ = -1} \alpha_{\mathbf{y}}$$

$$= \sum_{\mathbf{y} \in \mathcal{Y}} \alpha_{\mathbf{y}} \sum_{i=k-q}^{k+p} \delta(y_i^+ = -1) \overset{(*)}{=} \sum_{\mathbf{y} \in \mathcal{Y}} \alpha_{\mathbf{y}} (q+1) = q+1,$$

where equality (*) is by (27). ■

Algorithmically, to check if Property 11 is satisfiable notice $\hat{\theta}_k^+$ must be the root of

$$H_k(\theta) = \sum_{i=1}^{k} (\beta_i^+(\theta) - 1) + \sum_{i=k+1}^{n_+} \beta_i^+(\theta), \tag{28}$$

where $\theta \in [a_k^+ - \mu, a_{k+1}^+]$,

and $-\frac{1}{n} - \theta \in [a_{k+1}^-, a_k^- - \mu]$.

Note we changed open interval in the last range into closed interval because it lies on the boundary of cases and does give a valid solution.

Clearly $H_k(\theta)$ is monotonically decreasing in $\theta$. To fall in this case, the following conditions are necessary and sufficient. First the above domain of $\theta$ is not empty. Second $H_k(\theta)$ at the maximum of the domain is non-positive and $R_k(\theta)$ at the minimum of the domain is non-negative. Third $a^-_{k+1} \le a^-_k - \mu$.

Once the root $\theta^*$ of $H_k(\theta)$ is found, set $k+p$ to the greatest $i$ such that $a^+_i \ge \theta^*$, and set to $k-q$ the smallest $i$ such that $a^+_i - \mu \le \theta^*$. Assign $\hat\theta^+_i = \theta^*$ for all $k-q \le i \le k+p$. Then it is easy to see that $\sum_{i=k-q}^{k+p} \beta^+_i(\theta^*) = q+1$.

To find the root of $H_k(\theta)$ we only need to use binary search which costs $O(\log n)$ steps with each step costing $O(\log n)$ time for evaluating $H_k(\theta)$. This requires some pre-computation which is similar to the root finding procedure for $G(\theta)$ in case 1.2.

**Case 2.2.2.** $\hat\theta^+_{k+1} + \hat\theta^-_{k+1} + \frac{1}{n} = 0$. Now both $\hat\theta^+_k$ and $\hat\theta^-_{k+1}$ can be tied. Suppose $\hat\theta^+_{k+p+1} < \hat\theta^+_{k+p} = \ldots = \hat\theta^+_{k-q} < \hat\theta^+_{k-q-1}$ where $p, q \ge 0$, and $\hat\theta^-_{k+r+1} < \hat\theta^-_{k+r} = \ldots = \hat\theta^-_{k+1}$ where $r \ge 1$. Then using the same proof technique as above, we can show that for all $i > k+p$, $\hat\theta^+_i = a^+_i$ and for all $j > k+r$, $\hat\theta^-_j = a^-_j$. For all $i < k-q$, $\hat\theta^i = a^+_i - \mu$. So the only values to be determined are $\hat\theta^+_k (= \hat\theta^+_i, \ \forall\ k-q \le i \le k+p)$ and $\hat\theta^-_{k+1} (= \ldots = \hat\theta^-_{k+r})$. This can be accomplished by using the following property, which can be viewed as a combination of Property 8 and Property 11.

**Property 12.** $\sum_{i=k-q}^{k+p} \hat\beta^+_i = q+1+\sum_{j=k+1}^{k+r} \hat\beta^-_j$.

*Proof.* First we write out the explicit form of $\mathcal{Y} := \operatorname{argmax}_{\mathbf{y}} q(\mathbf{y}, \hat{\boldsymbol\theta})$:

$$\Bigg\{ \mathbf{y} : \sum_{i=k-q}^{k+p} \delta(y^+_i = -1) = q+1+\sum_{j=k+1}^{k+r} \delta(y^-_j = 1), \\ y^+_i = -1\ \forall i < k-q, \text{ and } y^+_i = 1\ \forall i > k+p, \\ y^-_j = 1\ \forall j \le k, \text{ and } y^-_j = -1\ \forall j > k+r \Bigg\},$$

where the first condition is because $b(\mathbf{y}) = c(\mathbf{y})$, while $\sum_{i=1}^{k-q-1} \delta(y^+_i = -1) + \sum_{i=k+p}^{n_+} \delta(y^+_i = -1) = k-q-1$ and $\sum_{j=1}^{k} \delta(y^-_j = 1) + \sum_{j=k+r+1}^{n_-} \delta(y^-_j = 1) = k$. Since $\mathbf{0} \in \partial D(\hat{\boldsymbol\theta})$, so there exists a distribution $\boldsymbol\alpha$ over $\mathbf{y} \in \mathcal{Y}$ such that for all $k-q \le i \le k+p$ and $k \le j \le k+r$:

$$\hat\beta^+_i = \sum_{\mathbf{y}\in\mathcal{Y}: y^+_i = -1} \alpha_{\mathbf{y}}, \quad \text{and} \quad \hat\beta^-_j = \sum_{\mathbf{y}\in\mathcal{Y}: y^-_j = 1} \alpha_{\mathbf{y}}.$$

So

$$\sum_{i=k-q}^{k+p} \hat\beta^+_i = \sum_{i=k-q}^{k+p} \sum_{\mathbf{y}\in\mathcal{Y}: y^+_i = -1} \alpha_{\mathbf{y}} = \sum_{\mathbf{y}\in\mathcal{Y}} \alpha_{\mathbf{y}} \sum_{i=k-q}^{k+p} \delta(y^+_i = -1)$$
$$= \sum_{\mathbf{y}\in\mathcal{Y}} \alpha_{\mathbf{y}} \left( q+1+\sum_{j=k+1}^{k+r} \delta(y^-_j = 1) \right)$$
$$= q+1+\sum_{\mathbf{y}\in\mathcal{Y}} \alpha_{\mathbf{y}} \sum_{j=k+1}^{k+r} \delta(y^-_j = -1)$$
$$= q+1+\sum_{j=k+1}^{k+r} \sum_{\mathbf{y}\in\mathcal{Y}: y^-_j = 1} \alpha_{\mathbf{y}} = q+1+\sum_{j=k+1}^{k+r} \hat\beta^-_j. \qquad \blacksquare$$

Algorithmically, to check if Property 12 is satisfiable we notice $\hat\theta^+_{k+1}$ must be the root of

$$R_k(\theta) = \sum_{i=1}^{k} (\beta^+_i(\theta) - 1) + \sum_{i=k+1}^{n_+} \beta^+_i(\theta) - \sum_{j=k+1}^{n_-} \beta^-_j \left( -\frac{1}{n} - \theta \right),$$
where $\theta \in [a^+_k - \mu, a^+_{k+1}]$,
and $-\frac{1}{n} - \theta \in [a^-_{k+1} - \mu, \min\{a^-_k - \mu, a^-_{k+1}\}]$.

Clearly $R_k(\theta)$ is monotonically decreasing in $\theta$, and has at most $6n$ linear pieces. To fall in this case, the following conditions are necessary and sufficient. First the above domain of $\theta$ is not empty. Second $R_k(\theta)$ at the maximum of the domain is non-positive and $R_k(\theta)$ at the minimum of the domain is non-negative.

After obtaining the root $\theta^*$, set $k+p$ to the greatest $i$ such that $a^+_i \ge \theta^*$, and set $k-q$ to the smallest $i$ such that $a^+_i - \mu \le \theta^*$. Set $k+r$ to the greatest $j$ such that $a^-_j \ge -\frac{1}{n} - \theta^*$. Assign $\hat\theta^+_i = \theta^*$ for all $k-q \le i \le k+p$ and $\hat\theta^-_j = -\frac{1}{n} - \theta^*$. Then it is easy to see that $\sum_{i=k-q}^{k+p} \beta^+_i(\theta^*) = q+1+\sum_{j=k+1}^{k+r} \beta^-_j(-\frac{1}{n} - \theta^*)$.

The binary search for the root takes $O(\log n)$ steps and each step costs $(\log n)$ using the same pre-computation and root finding procedure as for $G(\theta)$ in case 1.2.

**Case 3.** Suppose $k = \min\{n_+, n_-\}$. Assume $n_+ \ge n_-$. Then we must have $\hat\beta^-_j = 1$ and $\hat\theta^-_j = a^-_j - \mu$ for all $j \in [n_-]$. The proof is similar to the above and we omit it here. Suppose $\hat\theta^+_{k+p+1} < \hat\theta^+_{k+p} = \ldots = \hat\theta^+_k = \ldots = \hat\theta^+_{k-q} < \hat\theta^+_{k-q-1}$ where $p, q \ge 0$. Then we have $\hat\beta^+_i = 1$ and $\hat\theta^+_i = a^+_i - \mu$ for all $i < k-q$, and $\hat\beta^+_i = 0$ and $\hat\theta^+_i = a^+_i$ for all $i > k+p$. Similar to Property 11, we can also show that $\sum_{i=k-q}^{k+p} \hat\beta^+_i = q+1$, which can be used to determine $\hat\theta^+_i$ $(k-q \le i \le k+p)$. The root finding by binary search is also the same and the cost

is $O(\log n)$. Everything is the same as the $H_k$ in (28), except that we treat $a^-_{n_-+1} = -\infty$.

**Summary** Our algorithm is simple: enumerate $k$ from 0 to $\min\{n_+, n_-\}$ and check whether the condition of any of the above cases can be satisfied. Since we know there is a unique solution, hence there must exist a unique $k$ and a case which is satisfied. Since each check takes $O(\log n)$ time, the total time cost is $O(n \log n)$. In a more fancy fashion, we can even do a binary search over $k$, but it will require more refined characterization of whether to increase or decrease $k$ when none of the cases is satisfied for a given $k$.

## F Vishy's algorithm

**Claim 1:** There exists a $\theta^+$ and a $\theta^-$ such that the argmin of (20) can we written as $\theta^+_i = \theta^+$ for $i = 1, \ldots, n^+$ and $\theta^-_j = \theta^-$ for $j = 1, \ldots, n^-$.

**Claim 2:** The function

$$f^+(\theta) := \sum_{i=1}^{n_+} \nabla h_i(\theta) = \sum_{i=1}^{n_+} \nabla \beta^+_i(\theta) \tag{29}$$

is a piecewise linear non-decreasing function of $\theta$ with at most $2n_+$ kink points. The same holds for

$$f^-(\theta) := \sum_{j=1}^{n_-} \nabla h_j(\theta) = \sum_{j=1}^{n_-} \nabla \beta^-_j(\theta). \tag{30}$$

*Proof.* Since $h_i$ is a convex function of $\theta$ and therefore $\sum_i h_i$ is also a convex function, while $f^+$ is its gradient. Consequently, $f^+$ is non-decreasing. Furthermore, $\nabla h_i$ defined in (21) is a piecewise linear function with 2 kink points. Summing piecewise linear functions gives rise to a piecewise linear function with at most $2n^+$ kinks. An identical argument can be made for $f^-$. ■

**Claim 3:** At the optimal solution of (20)

$$\sum_{i=1}^{n_+} \nabla h_i(\theta) = \sum_{j=1}^{n_-} \nabla h_j(\theta).$$

In other words, $f^+(\theta^+) = f^-(\theta^-)$.

*Proof.* Essentially similar to the proof of property 4. ■

The high level description of the algorithm is as follows:

- Sort the arrays $\{a^+_i, a^+_i - \mu\}$ with $i = 1, \ldots, n^+$ and $\{a^-_j, a^-_j - \mu\}$ with $j = 1, \ldots, n^-$. I will call the entries of $\{a^+_i, a^+_i - \mu\}$ (resp. $\{a^-_j, a^-_j - \mu\}$) as $b^+_i$ (resp. $b^-_j$) below
- Evaluate $f^+(b^+_i)$ and $f^-(b^-_j)$. This takes $O(n)$ time given the sorted arrays
- Find $i$ and $j$ such that one of the two following conditions holds:

  $f^+(b^+_i) \le f^-(b^-_j)$ and $f^+(b^+_{i+1}) \ge f^-(b^-_{j+1})$ or
  $f^+(b^+_i) \ge f^-(b^-_j)$ and $f^+(b^+_{i+1}) \le f^-(b^-_{j+1})$

  Then $\theta^+$ lies in the interval $(b^+_i, b^+_{i+1})$ while $\theta^-$ lies in the interval $(b^-_j, b^-_{j+1})$.

**Algorithm 1** Compute $f^+$

**Require:** $a^+_0 \ge a^+_1 \ge \ldots \ge a^+_{n_+-1}$
**Require:** Smoothing parameter $\mu$
$i_1 = 0$, $i_2 = 0$, $i = 0$
$s = 0$ (slope)
$f_{-1} = 0$ (array)
**while** $i_1 < n^+$ and $i_2 < n^+$ **do**
  **if** $a^+_{i_1} > a^+_{i_2} - \mu$ **then**
    $s \leftarrow s + a^+_{i_1}$
    $f_i \leftarrow f_{i-1} + \frac{1}{\mu}(a^+_{i_1} - s)$
    $i_1 \leftarrow i_1 + 1$
  **else**
    $i_2 \leftarrow i_2 + 1$
    $p = a_{i_2} - \mu$
  **end if**
  $i \leftarrow i + 1$
**end while**

## G Plots of All Experimental Results

## H Plots with Nesterov's Solver

We tried to use FISTA (Beck & Teboulle, 2009) to optimize the smoothed objective. It turns out that it is quite effective when a high accuracy solution is need. However, it is slow initially.

We only used $\mu = 100\hat{\mu}$ because it is already slower than other solvers. In all the figures, the cyan line with triangle marker is the result of FISTA.