Robust Forward Algorithms via PAC-Bayes and Laplace Distributions

Asaf Noy
Supervisor: Prof. Koby Crammer

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, Israel

March 31th, 2014

Outline

1. Introduction
   - PAC-Bayes theory
   - Boosting
   - Robust Statistics

Outline

1. Introduction
   - PAC-Bayes theory
   - Boosting
   - Robust Statistics
2. Laplace-like family of distributions
   - Laplace-like Regularization
   - Laplace-like and PAC-Bayes bounds

Outline

1. Introduction
   - PAC-Bayes theory
   - Boosting
   - Robust Statistics
2. Laplace-like family of distributions
   - Laplace-like Regularization
   - Laplace-like and PAC-Bayes bounds
3. PAC-Bayesian Boosting Algorithms
   - The ExpLoss: Huber-Reg AdaBoost
   - The LogLoss: BaLaBoost

Preliminaries

- **Binary classification:** $x \in \mathcal{X} \subseteq \mathbb{R}^d$, $y \in \mathcal{Y} = \{\pm 1\}$, $h : \mathcal{X} \mapsto \mathcal{Y}$.
- **Zero-one loss:**

$$\ell_{zo}\left(y(\boldsymbol{\omega} \cdot \boldsymbol{x})\right) = \begin{cases} 1 & y(\boldsymbol{\omega} \cdot \boldsymbol{x}) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

- **Linear classifiers:** $h(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{\omega} \cdot \boldsymbol{x})$ for $\boldsymbol{\omega} \in \mathcal{H} \subseteq \mathbb{R}^d$.
- **Empirical risk:** $R_S(G_q) = \frac{1}{m} \sum_{i=1}^{m} \ell_{zo}(y_i(\boldsymbol{x}_i \cdot \boldsymbol{\omega}))$.
- **Joint distribution assumption:** $(\boldsymbol{x}, y) \overset{iid}{\sim} \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$,
- **Expected loss (risk):** $R(G_q) = \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell_{zo}(y(\boldsymbol{x} \cdot \boldsymbol{\omega}))\right]$.

PAC-Bayes theory

- Name derives from Bayes theorem: we assume a prior distribution over classifiers and then use Bayes rule to update the prior based on the likelihood of the data for each classifier.
- First version proved by McAllester (1999).
- Improved proof and bound due to Seeger (2002) with application to Gaussian processes.
- Application to SVMs by Langford and Shaw-Taylor (2002).

PAC-Bayes theory

- The PAC-Bayes theorem involves a class of classifiers $\mathcal{H}$ together with a prior distribution P and posterior Q over $\mathcal{H}$.
- The distribution P must be chosen before learning, but the bound holds for all choices of Q, hence Q does not need to be the classical Bayesian posterior.

PAC-Bayesian theorem [Seeger, 2002]

Fix an arbitrary D, arbitrary prior P, and confidence δ, then with probability at least $1-\delta$ over samples $S \sim D^m$, all posteriors Q satisfy,

$$\mathrm{D_{KL}}(R_S(G_q) \| R(G_q)) \leq \frac{\mathrm{D_{KL}}(Q \| P) + \ln(m + 1/\delta)}{m}$$

with $R_S(G_q)$ and $R(G_q)$ considered as Bernoulli distributions' success parameters on $\{0,1\}$.

Boosting

- Koh et al., 2007, Duchi and Singer 2009
- Random Classification Noise Defeats All Convex Potential Boosters

Huber function

- The Huber loss function: [Robust statistics, 1974]

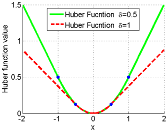
$$H_\delta(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le \delta \\ \delta(x - \frac{\delta}{2}) & |x| > \delta \end{cases}$$

- allows construction of an estimate which allows the effect of outliers to be reduced, while treating non-outliers in a more standard way.
- Often used in the context of robust filtering of Laplace-noise ([1],[2]).

[1] Robust estimation using the Huber function with a data-dependent tuning constant (You-Gan et al. , 2007).
[2] An l1-laplace robust kalman smoother (Aravkin et al. , 2011).

In this work

- Derive robust boosting-like algorithms directly from PAC-Bayes bounds, and analyze their properties.
- Generalize PAC-Bayes theory to the multi-task framework.
- Testing the algorithms in a wide range of input noise.



Huber Fucntion δ=0.5
Huber Fucntion δ=1
Huber function value
x
1.5
1
0.5
0
-2
-1
0
1
2

Laplace-like family of distributions

- Let $Q(\boldsymbol{\omega};\boldsymbol{\mu},\boldsymbol{\sigma}) \in \mathcal{L}^2$, then,

$$Q(\boldsymbol{\omega};\boldsymbol{\mu},\boldsymbol{\sigma}) = \frac{1}{2^d \prod_{k=1}^{d} \sigma_k} e^{-\|\boldsymbol{\omega}-\boldsymbol{\mu}\|_{\boldsymbol{\sigma},1}} \ , \ \|\boldsymbol{\omega}\|_{\boldsymbol{\sigma},1} = \frac{1}{2d} \sum_{k=1}^{d} \frac{|\omega_k|}{\sigma_k}$$

- Uni-modal distribution with mean $\boldsymbol{\mu}$, and diagonal covariance matrix $\Sigma = 2 \times \mathrm{diag}\left(\sigma_1^2, .., \sigma_d^2\right)$.
- **Proposition:** The single continuous d-dimensional distribution $Q(\boldsymbol{\omega})$ with a bounded expected $\boldsymbol{\sigma}$-weighted ℓ_1-norm, $E\left(\|\boldsymbol{\omega}-\boldsymbol{\mu}\|\right)_{\boldsymbol{\sigma},1} \leq 1$, which maximizes the information-theoretic entropy maintains $Q \in LL$.

Theorem 1

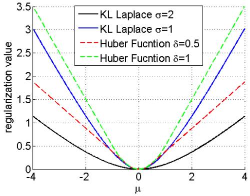
Let $P\left(\boldsymbol{\mu}_P, \boldsymbol{\sigma}_P\right), Q\left(\boldsymbol{\mu}_Q, \boldsymbol{\sigma}_Q\right) \in \mathcal{L}^2$ be two LL distributions. The KL-divergence between these two distributions is well defined and given by,

$$\mathrm{D_{KL}}(Q\|P) = \sum_{k=1}^{d} \left[\frac{\sigma_{Q,k}}{\sigma_{P,k}} \left(\frac{|\mu_{Q,k}-\mu_{P,k}|}{\sigma_{Q,k}} + e^{-\frac{|\mu_{Q,k}-\mu_{P,k}|}{\sigma_{Q,k}}} \right) + \log\left(\frac{\sigma_{P,k}}{\sigma_{Q,k}}\right) - 1 \right].$$

Laplace-like family of distributions

- Setting $\mu_P = 0$ and $\sigma_Q = \sigma_P$ in the 1-dimensional case, We obtain,

$$g_{\sigma_Q}(\mu_Q) = \frac{|\mu_Q|}{\sigma_Q} + \exp\left\{-\frac{|\mu_Q|}{\sigma_Q}\right\} - 1$$
$$\approx \begin{cases} \frac{|\mu_Q|}{\sigma_Q} - 1 & |x| \gg 1 \\ \frac{1}{2}\left(\frac{|\mu_Q|}{\sigma_Q}\right)^2 & |x| \ll 1 \end{cases}$$



1. Huber function $\notin C^2 \implies g_{\sigma_Q}(\mu_Q) \in C^\infty$.
2. Huber function is convex $\implies g_{\sigma_Q}(\mu_Q)$ is strictly-convex.
3. $g_{\sigma_Q}(\mu_Q)$ is closely related to the *linear loss* ℓ ... (later...)

Laplace-like and PAC-Bayes bounds

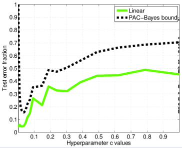
Theorem 2 (PAC-Bayes for linear classifiers)

For any distribution $\mathcal{D}$, any set $\mathcal{H}$ of classifiers, any distributions P, Q of support $\mathcal{H}$, any $\delta \in \{0, 1\}$, and any positive real scalar c, we have:

$$\mathrm{E}_{\boldsymbol{\omega}\sim Q,(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell_{zo}(y(\boldsymbol{x}\cdot\boldsymbol{\omega}))\right] \leq \frac{1}{1-\exp(-c)}\times$$

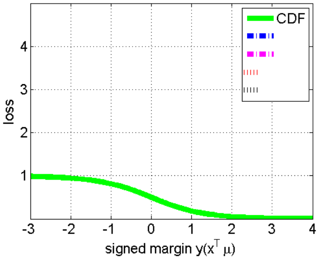
$$\left[1-\exp\left\{-\frac{1}{m}\left(cE_{\boldsymbol{\omega}\sim Q}\left[\sum_{i=1}^{m}\ell_{zo}(y_i(\boldsymbol{x}_i\cdot\boldsymbol{\omega}))\right]+\mathrm{D}_{\mathrm{KL}}(Q\|P)+\ln\frac{1}{\delta}\right)\right\}\right]$$

with probability of at least $1-\delta$.



Laplace-like and PAC-Bayes bounds

- The error term $\Pr\left(y_i(\boldsymbol{x}_i \cdot \boldsymbol{\omega}) \leq 0\right)$ is never convex for each $\boldsymbol{\mu}$.
- Two ways to go:
  1. Upper-bound the error term with smooth and convex functions.
  2. Directly calculate the error term, then bound the result with smooth and convex functions.
- As we shall see later- the tighter the bound, the better the results..



The ExpLoss

- **Assumptions:**
  1. Isotropic $\mathcal{L}^2$ distributions, $\forall k : \sigma_{Q,k} = \sigma$.
  2. Bounded input, $\max_{1 \leq i \leq m} \|\boldsymbol{x}_i\|_\infty < 1$.
- Consider the ExpLoss: $\ell_{exp}\left(y(\boldsymbol{\omega} \cdot \boldsymbol{x})\right) = \mathrm{E}_Q\left[e^{-y\boldsymbol{x}\cdot\boldsymbol{\omega}}\right]$.
- Define scaled mean vector, $\mu_k = \frac{\mu_{Q,k}}{\sigma}$. The resulting objective,

$$\mathcal{F}_{\exp}\left(\boldsymbol{\mu}, \sigma\right) = -d \log \sigma + \sigma \sum_{k=1}^{d} \left(|\mu_k| + e^{-|\mu_k|}\right) + c \sum_{i=1}^{m} D_i e^{-\sigma y_i \boldsymbol{x}_i \cdot \boldsymbol{\mu}},$$

$$\text{for } D_i = D_i\left(\boldsymbol{\sigma}_Q\right) = \prod_{k=1}^{d} \left(1 - \left(x_{i,k}\sigma_{Q,k}\right)^2\right)^{-1}.$$

- For a fixed σ, we perform coordinate descent over $\boldsymbol{\mu}$.

The ExpLoss

- Define $C_{+}(\boldsymbol{\mu}^{(k)},\sigma) = c\sum_{i=1}^{m} D_i e^{-\sigma y_i \boldsymbol{x}_i^{(k)}\cdot\boldsymbol{\mu}^{(k)}}\left(\frac{1+\sigma y_i x_{ik}}{2}\right)$
  $C_{-}(\boldsymbol{\mu}^{(k)},\sigma) = c\sum_{i=1}^{m} D_i e^{-\sigma y_i \boldsymbol{x}_i^{(k)}\cdot\boldsymbol{\mu}^{(k)}}\left(\frac{1-\sigma y_i x_{ik}}{2}\right)$
- For the non-regularized objective we would get, $\mu_k = 0.5\ln\left(\frac{C_+}{C_-}\right)$.

The Huber-Reg AdaBoost algorithm

- **Input:** Train set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$, $\boldsymbol{\mu}_P \in \mathbb{R}^d$, $\boldsymbol{\sigma} \in (0,1)$, $c > 0$, $T > 0$.
- **Initialization**: $\boldsymbol{\mu}_Q^{(1)} = \boldsymbol{\mu}_P$; D_i for $i = 1, \ldots, m$.
  Loop For $t = 1, \ldots, T$ do:
  - Choose coordinate: $k \in \{1, .., d\}$, and set: $C_{+}(\boldsymbol{\mu}^{(k)},\sigma)$, $C_{-}(\boldsymbol{\mu}^{(k)},\sigma)$.
  - Update: if $(C_+ \geq C_-)$
    then $\mu_{Q,k}^{(t+1)} \leftarrow \mu_{Q,k}^{(t)} + \log\left(\frac{-\sigma+\sqrt{\sigma^2+4C_-(\sigma+C_+)}}{2C_-}\right)$
    else $\mu_{Q,k}^{(t+1)} \leftarrow \mu_{Q,k}^{(t)} + \log\left(\frac{-\sigma+\sqrt{\sigma^2-4C_-(\sigma-C_+)}}{2(\sigma-C_+)}\right)$
- **Output:** $\boldsymbol{\mu}_Q^{(T+1)}$

The LogLoss

- Consider the LogLoss: $\ell_{log}\left(y(\boldsymbol{\omega}\cdot\boldsymbol{x})\right)=\log_2\left(1+\exp\left(-y\left(\boldsymbol{\omega}\cdot\boldsymbol{x}\right)\right)\right)$,

$$\mathcal{F}_{\log}\left(\boldsymbol{\mu}_Q,\boldsymbol{\sigma}_Q\right)=\sum_{k=1}^{d}\left(\sigma_{Q,k}e^{-\frac{\left|\mu_{Q,k}\right|}{\sigma_{Q,k}}}-\log\left(\sigma_{Q,k}\right)\right)+\|\boldsymbol{\mu}_Q\|_1+c\sum_{i=1}^{m}\log_2\left(1+D_i e^{-yx\cdots}\right.$$

- Define the incremental change $\delta_k^{(t)}=\mu_{Q,k}^{(t+1)}-\mu_{Q,k}^{(t)}$.

Theorem 3

The difference between the LogLoss objective evaluated at time t and time $t+1$ is lower bounded,

$$\mathcal{F}_{\log}(\boldsymbol{\mu}_Q^{(t)})-\mathcal{F}_{\log}(\boldsymbol{\mu}_Q^{(t+1)})\geq c\sigma_{Q,k}\left(\gamma_k^{+}\left[1-e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right]+\gamma_k^{-}\left[1-e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right]\right)$$

$$+\left|\mu_{Q,k}^{(t)}\right|+\sigma_{Q,k}e^{-\frac{\left|\mu_{Q,k}^{(t)}\right|}{\sigma_{Q,k}}}-\left|\mu_{Q,k}^{(t)}+\delta_k^{(t)}\right|-\sigma_{Q,k}e^{-\frac{\left|\mu_{Q,k}^{(t)}+\delta_k^{(t)}\right|}{\sigma_{Q,k}}}.$$

$$q_t(i)=D_i/\left(D_i+e^{y_i\boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}\right),\ \gamma_k^{\pm}=\sum_{i=1}\mathbf{1}(y_i x_{i,k}\in\pm)q_t(i)\left|x_{i,k}\right|.$$

The LogLoss

- omitting terms independent of $\delta_k^{(t)}$, we can minimize the following,

$$\arg\min_{\delta_k^{(t)}} \left[\left|\mu_{Q,k}^{(t)} + \delta_k^{(t)}\right| + \sigma_{Q,k} e^{-\frac{\left|\mu_{Q,k}^{(t)}+\delta_k^{(t)}\right|}{\sigma_{Q,k}}} + c\sigma_{Q,k} \left(\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} + \gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right.\right.$$

BaLaBoost algorithm [Noy and Crammer, 2014]

- **Input:** Train set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, $\boldsymbol{\mu}_P \in \mathbb{R}^d$, $\boldsymbol{\sigma}_Q \in (0,1)^d$, $c > 0$, $T > 0$.
- **Initialization:** $\boldsymbol{\mu}_Q^{(1)} = \boldsymbol{\mu}_P$; D_i for $i = 1, \ldots, m$.
  Loop For $t = 1, \ldots, T$ do:
  - Choose coordinate: $k \in \{1, .., d\}$, and set: γ_k^+ , γ_k^-.
  - Update: if $\left(\gamma_k^+ \exp\left\{\frac{2\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}\right\} \geq \gamma_k^-\right)$

    then $\mu_{Q,k}^{(t+1)} \leftarrow \mu_{Q,k}^{(t)} + \delta_{k,+}^{(t)}(\gamma_k^+, \gamma_k^-, \mu_{Q,k}, \sigma_{Q,k})$

    else $\mu_{Q,k}^{(t+1)} \leftarrow \mu_{Q,k}^{(t)} + \delta_{k,-}^{(t)}(\gamma_k^+, \gamma_k^-, \mu_{Q,k}, \sigma_{Q,k})$
- **Output:** $\boldsymbol{\mu}_Q^{(T+1)}$

The LogLoss

- omitting terms independent of $\delta_k^{(t)}$, we can minimize the following,

$$\arg\min_{\delta_k^{(t)}} \left[\left|\mu_{Q,k}^{(t)} + \delta_k^{(t)}\right| + \sigma_{Q,k} e^{-\frac{\left|\mu_{Q,k}^{(t)}+\delta_k^{(t)}\right|}{\sigma_{Q,k}}} + c\sigma_{Q,k} \left(\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} + \gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right.\right.$$

BaLaBoost algorithm [Noy and Crammer, 2014]

- **Input:** Training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, $\boldsymbol{\mu}_P \in \mathbb{R}^d$, $\boldsymbol{\sigma}_Q \in (0,1)^d$, $c > 0$, No. of iterations T.
- **Initialization**: $\boldsymbol{\mu}_Q^{(1)} = \boldsymbol{\mu}_P \quad ; \quad D_i \quad \text{for } i = 1, \ldots, m$

  Loop For $t = 1, \ldots, T$ do:
  - Choose coordinate: $k \in \{1, .., d\}$, and set: γ_k^+ , γ_k^-.
  - Update: If $\left(\gamma_k^+ \exp\left\{\frac{2\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}\right\} \geq \gamma_k^-\right)$

    then $\mu_{Q,k}^{(t+1)} \leftarrow \mu_{Q,k}^{(t)} + \sigma_{Q,k} \log \left(\frac{1+\sqrt{1+4c\gamma_k^-\left[\exp\left\{-\frac{\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}\right\}+c\gamma_k^+\right]}}{2c\gamma_k^-} \right.$

The LogLoss

- g