Margins, Kernels and Non-linear Smoothed Perceptrons

Carnegie Mellon University

Aadtiya Ramdas and Javier Peña

Normalized Margins and Perceptron Algorithm

Non-linear Feasibility Problems

Given n points $x_1, \dots, x_n \in \mathbb{R}^d$
and labels $y_1, \dots, y_n \in \{-1, +1\}$

Goal: find unit vector $w \in \mathbb{R}^d$ s.t.

$$y_i(w^T x_i) \geq 0 \ \text{ i.e. } \ \text{sign}(w^T x_i) = y_i$$

Nonlinear Goal: find unit norm function $f \in \mathcal{F}_K$ s.t.

$$y_i f(x_i) \geq 0 \ \text{ i.e. } \ \text{sign}(f(x_i)) = y_i$$

Builds heavily on work by Negar Soheili + Javier Peña '12,'13

Unnormalized Margin

Nonlinear Goal: find unit norm function $f \in \mathcal{F}_K$ s.t.

$$y_i f(x_i) \geq 0 \ \text{ i.e. } \ \text{sign}(f(x_i)) = y_i$$

This has unnormalized margin $\rho > 0$ if $\exists f$ s.t.

$$y_i f(x_i) \geq \rho$$

or correspondingly in the linear case,

$$y_i w^T x_i \geq \rho$$

Normalized > Unnormalized Margin

Denote $X_2 = [x_1/\|x_1\|_2, \dots, x_n/\|x_n\|]$.

Define $\rho := \max_{\|w\|_2=1} \min_{p \in \Delta_n} \langle YX^T w, p \rangle$

$\rho_2 := \max_{\|w\|_2=1} \min_{p \in \Delta_n} \langle YX_2^T w, p \rangle$

where $Y = \text{diag}(y), X = [x_1, \dots, x_n]$.

...

Then $\dfrac{\rho}{\max_i \|x_i\|_2} \leq \rho_2$

Simple example given in the paper.

Normalized Perceptron

Algorithm 2 Normalized Perceptron

Initialize $w_0 = 0, p_0 = 0$
for $k = 0, 1, 2, 3, \dots$ **do**
  if $YX^\top w_k > 0$ **then**
    Exit, with w_k as solution
  else
    $\theta_k := \frac{1}{k+1}$
    $w_{k+1} := (1 - \theta_k) w_k + \theta_k XYp(w_k)$
  end if
end for

$$p(w) = \arg\min_{p \in \Delta_n} \langle YX^T w, p \rangle$$

where Δ_n is the n-dimensional probability simplex.

If $\rho_2 > 0$, then it finds a perfect separator in $\frac{1}{\rho_2^2}$ iterations.

Smoothed Normalized Kernel Perceptron (NKP)

Normalized (Kernel) Margin

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a psd kernel, giving rise to RKHS $\mathcal{F}_K$.

At each $x \in \mathbb{R}^d$, let $\phi_x : \mathbb{R}^d \to \mathbb{R}$ be the associated feature map, where $\phi_x(y) = K(x, y)$ and inner product $\langle \phi_x, \phi_y \rangle_K = K(u, v)$.

Define the normalized feature map

$$\tilde{\phi}_x = \frac{\phi_x}{\sqrt{K(x,x)}} \in \mathcal{F}_K \ \text{ and } \ \tilde{\phi}_X = [\tilde{\phi}_{x_1}, \dots, \tilde{\phi}_{x_n}].$$

We use the notation

$$Y\tilde{f}(X) = \left[y_i \frac{f(x_i)}{\sqrt{K(x_i, x_i)}} \right]_{i=1}^n.$$

Finally, the normalized margin is defined as

$$\rho_K := \sup_{\|f\|_K = 1} \inf_{p \in \Delta_n} \langle Y\tilde{f}(X), p \rangle.$$

Normalized Kernel Perceptron

Algorithm 3 Normalized Kernel Perceptron (NKP)

Set $\alpha_0 := 0$
for $k = 0, 1, 2, 3, \dots$ **do**
  if $G\alpha_k > \mathbf{0}_n$ **then**
    Exit, with α_k as solution
  else
    $\theta_k := \frac{1}{k+1}$
    $\alpha_{k+1} := (1 - \theta_k)\alpha_k + \theta_k p(\alpha_k)$
  end if
end for

$G_{ji} = G_{ij} := \frac{y_i y_j K(x_i, x_j)}{\sqrt{K(x_i,x_i)K(x_j,x_j)}} = \langle y_i \tilde{\phi}_{x_i}, y_j \tilde{\phi}_{x_j} \rangle_K$, and

$$p(\alpha) := \arg\min_{p \in \Delta_n} \langle \alpha, p \rangle_G, \ \langle p, \alpha \rangle_G := p^\top G \alpha, \ \|\alpha\|_G := \sqrt{\alpha^\top G \alpha}$$

If $\rho_K > 0$, then it finds a perfect separator in $\frac{1}{\rho_K^2}$ iterations.

NKP turns out to be a subgradient algorithm for minimizing

$$L(f) = \left\{ \sup_{p \in \Delta_n} \left\langle -Y\tilde{f}(X), p \right\rangle \right\} + \tfrac{1}{2}\|f\|_K^2.$$

By Representer theorem, $f^* = \sum_i \alpha_i y_i \phi_{x_i}$, so we can consider

$$L(\alpha) := \left\{ \sup_{p \in \Delta_n} \langle -\alpha, p \rangle_G \right\} + \frac{1}{2}\|\alpha\|_G^2$$

Lemma 1. $L(\alpha) < 0$ implies $G\alpha > 0$ and there exists a perfect classifier iff $G\alpha > 0$.

Lemma 2. For any $\alpha \in \mathbb{R}^n$, $\|\alpha\|_G \leq \|\alpha\|_1 \leq \sqrt{n}\|\alpha\|_2$.

Lemma 3. When $\rho_K > 0$, f maximizes the margin iff $\rho_K f$ optimizes $L(f)$. Hence, the margin is equivalently

$$\rho_K = \sup_{\|\alpha\|_G = 1} \inf_{p \in \Delta_n} \langle \alpha, p \rangle_G \leq \|p\|_G \quad \text{for all } p \in \Delta_n.$$

Smoothed NKP

Algorithm 4 Smoothed Normalized Kernel Perceptron

Set $\alpha_0 = \mathbf{1}_n/n, \ \mu_0 := 2, \ p_0 := p_{\mu_0}(\alpha_0)$
for $k = 0, 1, 2, 3, \dots$ **do**
  if $G\alpha_k > \mathbf{0}_n$ **then**
    Halt: α_k is solution to Eq. (8)
  else
    $\theta_k := \frac{2}{k+3}$
    $\alpha_{k+1} := (1 - \theta_k)(\alpha_k + \theta_k p_k) + \theta_k^2 p_{\mu_k}(\alpha_k)$
    $\mu_{k+1} = (1 - \theta_k)\mu_k$
    $p_{k+1} := (1 - \theta_k) p_k + \theta_k p_{\mu_{k+1}}(\alpha_{k+1})$
  end if
end for

$$p_\mu(\alpha) \ := \ \arg\min_{p \in \Delta_n} \left\{ \langle \alpha, p \rangle_G + \mu d(p) \right\} \ = \ \frac{e^{-G\alpha/\mu}}{\|e^{-G\alpha/\mu}\|_1},$$

where $\quad d(p) \ := \ \sum_i p_i \log p_i + \log n$

$$L_\mu(\alpha) = \sup_{p \in \Delta_n} \left\{ -\langle \alpha, p \rangle_G - \mu d(p) \right\} + \tfrac{1}{2}\|\alpha\|_G^2.$$

Lemma 4. (Lower Bound) At any step k, we have

$$L_{\mu_k}(\alpha_k) \geq L(\alpha_k) - \mu_k \log n.$$

Lemma 5. (Upper Bound) In any round k, SNKP satisfies

$$L_{\mu_k}(\alpha_k) \leq -\frac{1}{2}\|p_k\|_G^2.$$

Theorem 1. The SNKP algorithm finds a perfect classifier $f \in \mathcal{F}_K$ when one exists in $O\left(\frac{\sqrt{\log n}}{\rho_K}\right)$ iterations.

Von-Neumann (VN) and Gordan's Theorem

Gordan's Theorem

Exactly one of the following two statements can be true

1. Either there exists a $w \in \mathbb{R}^d$ such that for all i,
   $y_i(w^T x_i) > 0$,
2. Or, there exists a $p \in \Delta_n$ such that
   $\|XYp\|_2 = 0$, or equivalently $\sum_i p_i y_i x_i = 0$.

Von-Neumann-Gilbert Algorithm

Algorithm 5 Normalized Von-Neumann (NVN)

Initialize $p_0 = \mathbf{1}_n/n, w_0 = XYp_0$
for $k = 0, 1, 2, 3, \dots$ **do**
  if $\|XYp_k\|_2 \leq \epsilon$ **then**
    Exit and return p_k as an ϵ-solution to (13)
  else
    $j := \arg\min_i y_i x_i^\top w_k$
    $\theta_k := \arg\min_{\lambda \in [0,1]} \|(1 - \lambda) w_k + \lambda y_j x_j\|_2$
    $p_{k+1} := (1 - \theta_k) p_k + \theta_k e_j$
    $w_{k+1} := XYp_{k+1} = (1 - \theta_k) w_k + \theta_k y_j x_j$
  end if
end for

Von-Neumann described this algorithm in private communication with Dantzig in 1948, who then analyzed it but only published his proof in 1992. Independently, Gilbert created his algorithm in 1966.

1. When (D) is feasible, Von-Neumann-Gibert finds ϵ-certificate in $1/\epsilon^2$ steps.
2. Von-Neumann-Gilbert is a Frank-Wolfe method for: $\min_{p \in \Delta} \|Ap\|^2$
3. When (P) is feasible, Von-Neumann-Gilbert finds satisfying w in $1/\rho_A^{+2}$ steps.
4. When (D) is feasible, Von-Neumann-Gilbert finds ϵ-certificate in $\log(1/\epsilon)/|\rho_A^-|^2$ steps.

Dantzig (1992) proved (1), Nesterov verbally mentioned (2) to Epelman & Freund (1997) who proved (3,4).

Gordan's Theorem in RKHSs

Theorem 2 Exactly one of the following has a solution:

1. Either $\exists f \in \mathcal{F}_K$ such that for all i,
   $\frac{y_i f(x_i)}{\sqrt{K(x_i, x_i)}} = \langle f, y_i \tilde{\phi}_{x_i} \rangle_K > 0 \ \text{ i.e. } \ G\alpha > 0,$
2. Or $\exists p \in \Delta_n$ such that
   $\sum_i p_i y_i \tilde{\phi}_{x_i} = 0 \in \mathcal{F}_K \ \text{ i.e. } \ \|p\|_G = 0.$

Let us define the *witness set* as
$W := \{p \in \Delta_n | \sum_i p_i y_i \tilde{\phi}_{x_i} = 0\} = \{p \in \Delta_n | \|p\|_G = 0\}$

A Hoffman-bound for the dual

Lemma 7. For all $q \in \Delta_n$, the distance to the witness set

$$\text{dist}(q, W) := \min_{w \in W} \|q - w\|_2 \leq \min \left\{ \sqrt{2}, \frac{\sqrt{2}\|q\|_G}{|\rho_K|} \right\}.$$

As a consequence, $\|p\|_G = 0$ iff $p \in W$.

Theorem 3. When the primal is infeasible, the margin is

$$|\rho_K| = \sup \left\{ \delta \ \middle| \ \|f\|_K \leq \delta \implies f \in \text{conv}(Y\tilde{\phi}_X) \right\}$$

This quantity can be zero simply because an infinite dimensional ball cannot fit inside a finite dimensional hull. The *right* correction is to re-define the margin so that the only allowed w, f is in the affine hull of the points. Then, α can be used in Theorem 3 (for the "affine-margin", which can be non-zero even when the margin is zero).

Primal-dual Iterated Smoothed NKP-VN

Smoothed NKP-VN

Algorithm 6 Smoothed Normalized Kernel Perceptron-VonNeumann ($SNKPVN(q, \delta)$)

input $q \in \Delta_n$, accuracy $\delta > 0$
  Set $\alpha_0 = q, \ \mu_0 := 2n, \ p_0 := p_{\mu_0}^q(\alpha_0)$
  for $k = 0, 1, 2, 3, \dots$ **do**
    if $G\alpha_k > \mathbf{0}_n$ **then**
      Halt: α_k is solution to Eq. (8)
    else if $\|p_k\|_G < \delta$ **then**
      Return p_k
    else
      $\theta_k := \frac{2}{k+3}$
      $\alpha_{k+1} := (1 - \theta_k)(\alpha_k + \theta_k p_k) + \theta_k^2 \, p_{\mu_k}^q(\alpha_k)$
      $\mu_{k+1} = (1 - \theta_k)\mu_k$
      $p_{k+1} := (1 - \theta_k) p_k + \theta_k \, p_{\mu_{k+1}}^q(\alpha_{k+1})$
    end if
  end for

$$d_q^p = \tfrac{1}{2}\|p - q\|_2^2$$

$$p_\mu^q(\alpha) = \arg\min_{p \in \Delta_n} \langle G\alpha, p \rangle + \mu d_q^p,$$

$$L_\mu^q(\alpha) = \sup_{p \in \Delta_n} \left\{ -\langle \alpha, p \rangle_G - \mu d_q(p) \right\} + \tfrac{1}{2}\|\alpha\|_G^2$$

Lemma 8. [When $\rho_K > 0$ and $\delta < \rho_K$] For any $q \in \Delta_n$,

$$-\tfrac{1}{2}\|p_k\|_G^2 \ \geq \ L_{\mu_k}^q(\alpha_k) \ \geq \ L(\alpha_k) - \ \mu_k.$$

Hence SNKPVN finds a separator f in $O\left(\frac{\sqrt{n}}{\rho_K}\right)$ iterations.

Lemma 9. [When $\rho_K < 0$ or $\delta > \rho_K$] For any $q \in \Delta_n$,

$$-\tfrac{1}{2}\|p_k\|_G^2 \ \geq \ L_{\mu_k}^q(\alpha_k) \ \geq \ -\tfrac{1}{2}\mu_k \text{dist}(q, W)^2.$$

Hence SNKPVN finds a δ-solution in at most $O\left(\min\left\{\frac{\sqrt{n}}{\delta}, \frac{\sqrt{n}\|q\|_G}{\delta|\rho_K|}\right\}\right)$ iterations.

Typically we would be happy - we have a primal-dual algorithm! However, if we want the algorithm to have a *linear* convergence in δ, then we need to iterate it recursively as follows.

Iterated Smoothed NKP-VN

Algorithm 7 Iterated Smoothed Normalized Kernel Perceptron-VonNeumann ($ISNKPVN(\gamma, \epsilon)$)

input Constant $\gamma > 1$, accuracy $\epsilon > 0$
  Set $q_0 := \mathbf{1}_n/n$
  for $t = 0, 1, 2, 3, \dots$ **do**
    $\delta_t := \|q_t\|_G/\gamma$
    $q_{t+1} := SNKPVN(q_t, \delta_t)$
    if $\delta_t < \epsilon$ **then**
      Halt; q_{t+1} is a solution to Eq. (14)
    end if
  end for

Algorithm ISNKPVN satisfies

1. If the primal is feasible and $\epsilon < \rho_K$, then each call to SNKPVN halts in at most $\frac{2\sqrt{2n}}{\rho_K}$ iterations. Algorithm ISNKPVN finds a solution in at most $\frac{\log(1/\rho_K)}{\log(\gamma)}$ outer loops, bounding the total iterations by
$$O\left(\frac{\sqrt{n}}{\rho_K} \log\left(\frac{1}{\rho_K}\right)\right).$$
2. If the dual is feasible or $\epsilon > \rho_K$, then each call to SNKPVN halts in at most $O\left(\min\left\{\frac{\sqrt{n}}{\epsilon}, \frac{\sqrt{n}}{|\rho_K|}\right\}\right)$ steps. Algorithm ISNKPVN finds an ϵ-solution in at most $\frac{\log(1/\epsilon)}{\log(\gamma)}$ outer loops, bounding the total iterations by
$$O\left(\min\left\{\frac{\sqrt{n}}{\epsilon}, \frac{\sqrt{n}}{|\rho_K|}\right\} \log\left(\frac{1}{\epsilon}\right)\right).$$

It was unclear to us whether the $\sqrt{n}$ can be made $\sqrt{\log n}$ while

1. The algorithm still visually looks like the perceptron.
2. The algorithm achieves linear convergence w.r.t ϵ (for the dual)