SVMs + Margin bounds

CPSC 532S: Modern Statistical Learning Theory
2 March 2022
cs.ubc.ca/~dsuth/532S/22/

Admin

- Send in A1 regrade requests soon if I messed up
  - Either via Gradescope, or on Piazza

- A2 solutions are posted

- Later this week:
  - A3 (due in ~2 weeks after posting it)
  - Project info (with a short proposal also due in a few weeks)
  - If you don’t have a group and want one, post on Piazza asap

Linear classifiers

- Reminder: linear classifiers are
  - General case: $h(x) = \mathbb{I}(w^\top x + b \geq 0)$ for $x \in \mathbb{R}^d$
  - Homogenous: $h(x) = \mathbb{I}(w^\top x \geq 0)$

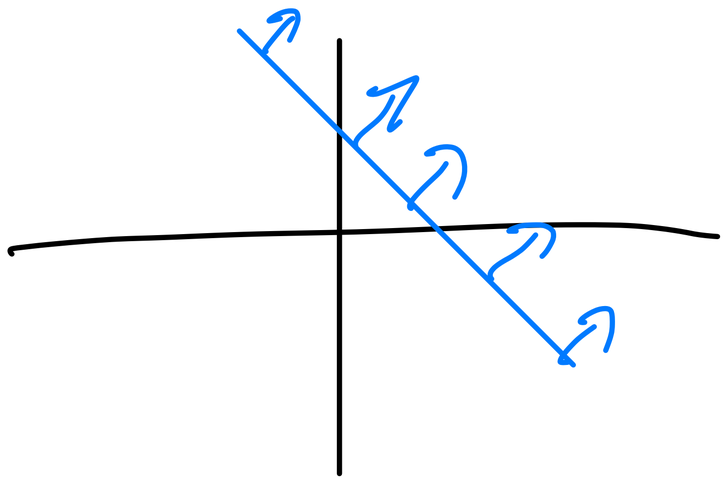
Linear classifiers

- Reminder: linear classifiers are
  - General case: $h(x) = \mathbb{I}(w^\top x + b \geq 0)$ for $x \in \mathbb{R}^d$
  - Homogenous: $h(x) = \mathbb{I}(w^\top x \geq 0)$
  - Can turn general case into homogenous by using $\tilde{x} = [1 \quad x] \in \mathbb{R}^{d+1}$

Linear classifiers

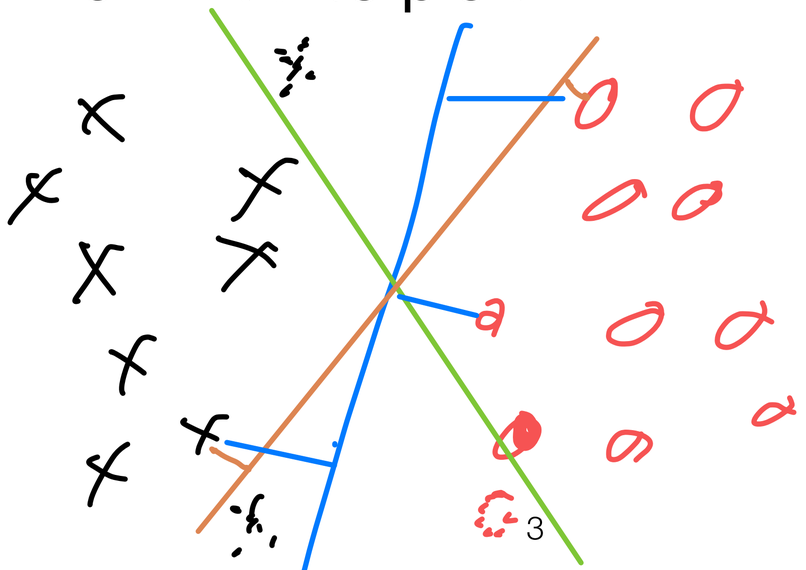
$h(x): x \to \{0,1\}$ $\quad x \to \{-1,1\}$ $\quad h(x) = \text{sign}(w^T x)$

- Reminder: linear classifiers are
  - General case: $h(x) = \mathbb{I}(w^\top x + b \geq 0)$ for $x \in \mathbb{R}^d$
  - Homogenous: $h(x) = \mathbb{I}(w^\top x \geq 0)$
  - Can turn general case into homogenous by using $\tilde{x} = [1 \quad x] \in \mathbb{R}^{d+1}$
  - Also called *halfspace, d*efined by a hyperplane $w^\top x + b = 0$



Linear classifiers

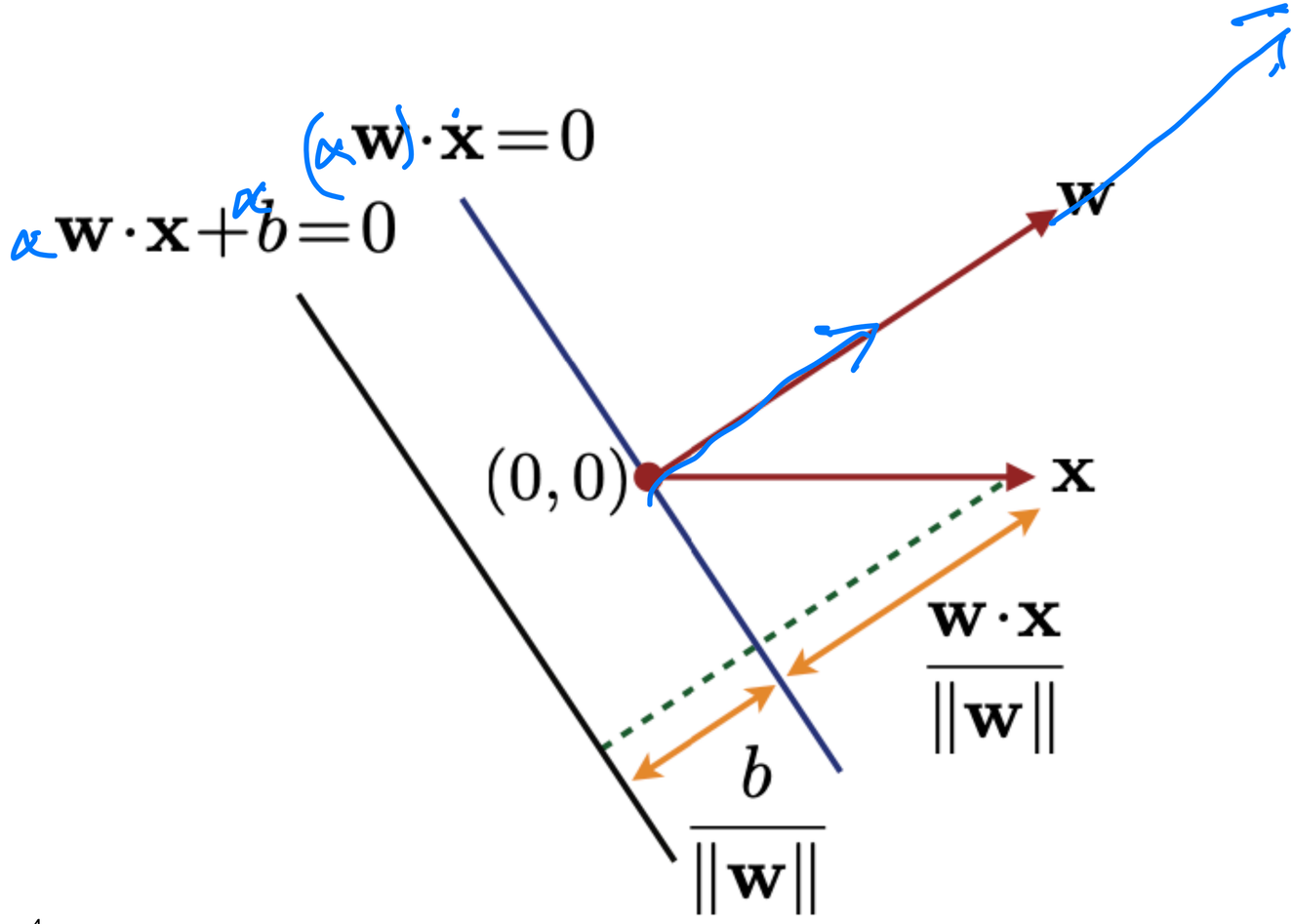
- Reminder: linear classifiers are
  - General case: $h(x) = \mathbb{I}(w^\top x + b \geq 0)$ for $x \in \mathbb{R}^d$
  - Homogenous: $h(x) = \mathbb{I}(w^\top x \geq 0)$
  - Can turn general case into homogenous by using $\tilde{x} = [1 \quad x] \in \mathbb{R}^{d+1}$
  - Also called *halfspace, d*efined by a hyperplane $w^\top x + b = 0$
- Separable case: which ERM to pick?



Margin

- Distance from x to hyperplane $\{x : w^\top x + b = 0\}$ is the *geometric margin*

$$\rho_h(x) = \frac{|w^\top x + b|}{\|w\|_2}$$

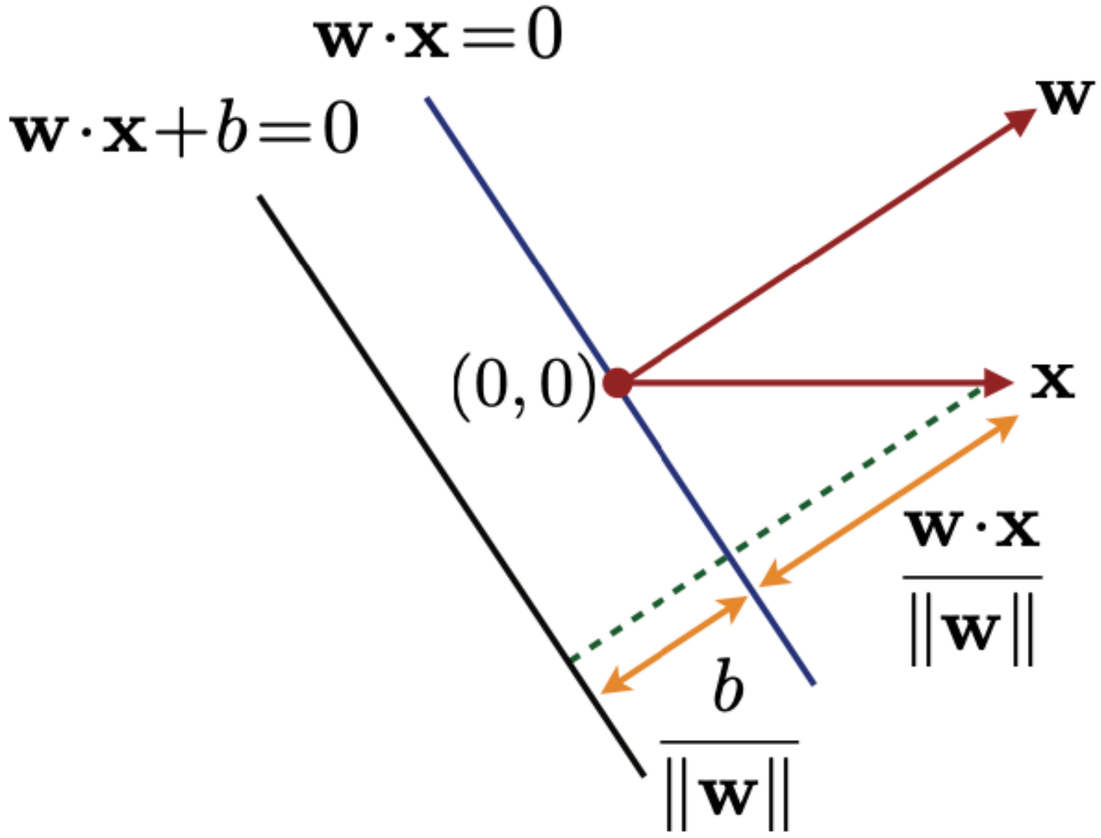


Margin

- Distance from x to hyperplane $\{x : w^\top x + b = 0\}$ is the *geometric margin*

$$\rho_h(x) = \frac{|w^\top x + b|}{\|w\|_2}$$

- Rescaling w and b doesn't change the classifier, or the margin



Max-margin classifier

- Margin for a training set is $\frac{1}{\|w\|} \min_{i \in [n]} |w^\top x_i + b|$

Max-margin classifier

- Margin for a training set is $\frac{1}{\|w\|} \min_{i \in [n]} |w^\top x_i + b|$
- Max-margin classifier (the **"hard" Support Vector Machine**) is

$$\arg\max_{w,b} \frac{1}{\|w\|} \min_{i \in [n]} |w^\top x_i + b| \quad \text{s.t.} \quad \forall i, \; y_i(w^\top x_i + b) > 0$$

Max-margin classifier

Geometric

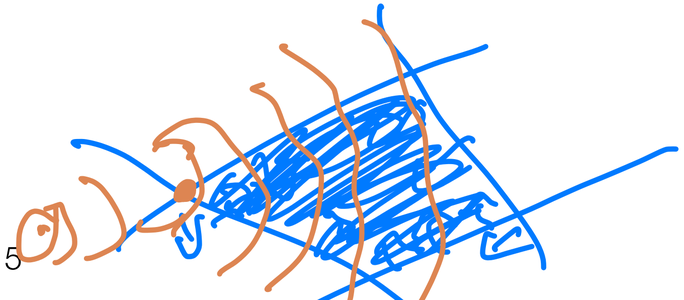
- Margin for a training set is $\frac{1}{\|w\|} \min_{i \in [n]} |w^\top x_i + b|$

- Max-margin classifier (the **"hard" Support Vector Machine**) is

$$\arg\max_{w,b} \frac{1}{\|w\|} \min_{i \in [n]} |w^\top x_i + b| \quad \text{s.t.} \quad \forall i, \; y_i(w^\top x_i + b) > 0$$

- Equivalently, as a **quadratic program**:

$$\arg\min_{w,b} \|w\|^2 \quad \text{s.t.} \quad \forall i, \; y_i(w^\top x_i + b) \geq 1$$

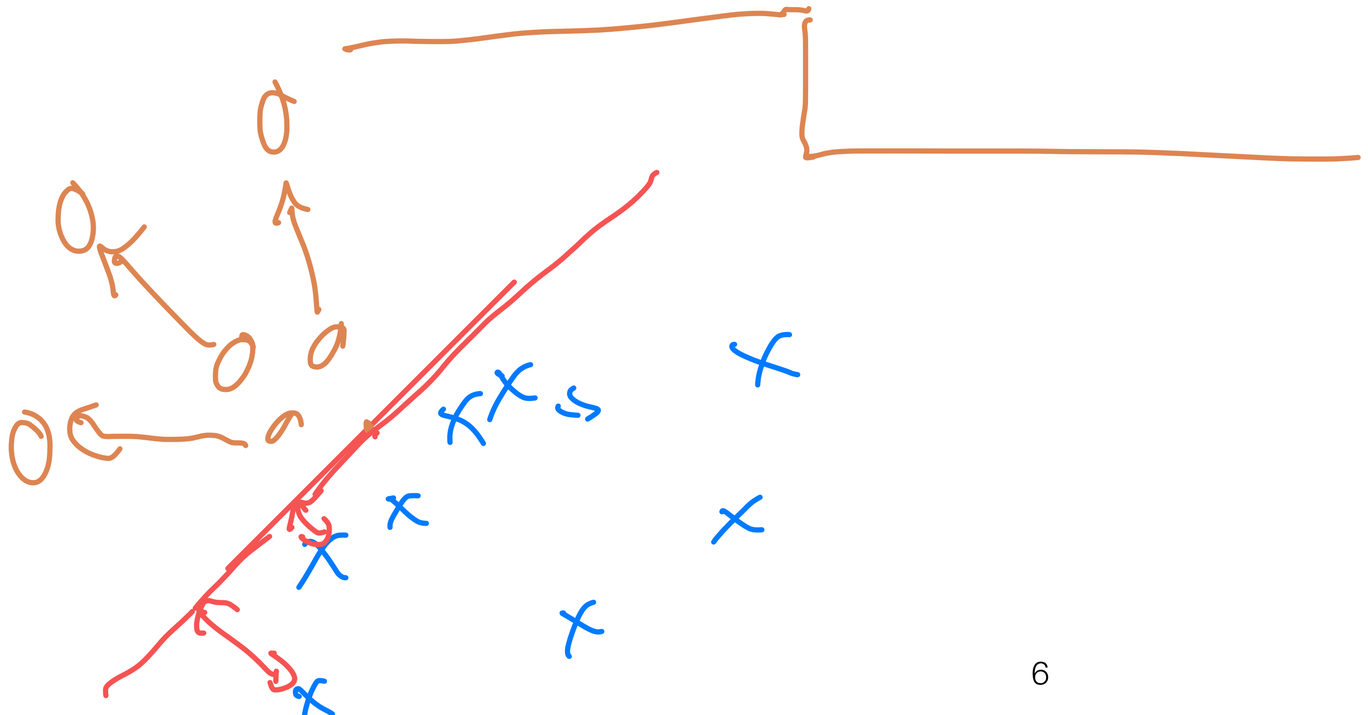


Margin-based bounds

- Would like to analyze 0-1 loss, but it’s not Lipschitz, so we can’t

Margin-based bounds

- Would like to analyze 0-1 loss, but it's not Lipschitz, so we can't
  - Not just a technicality!

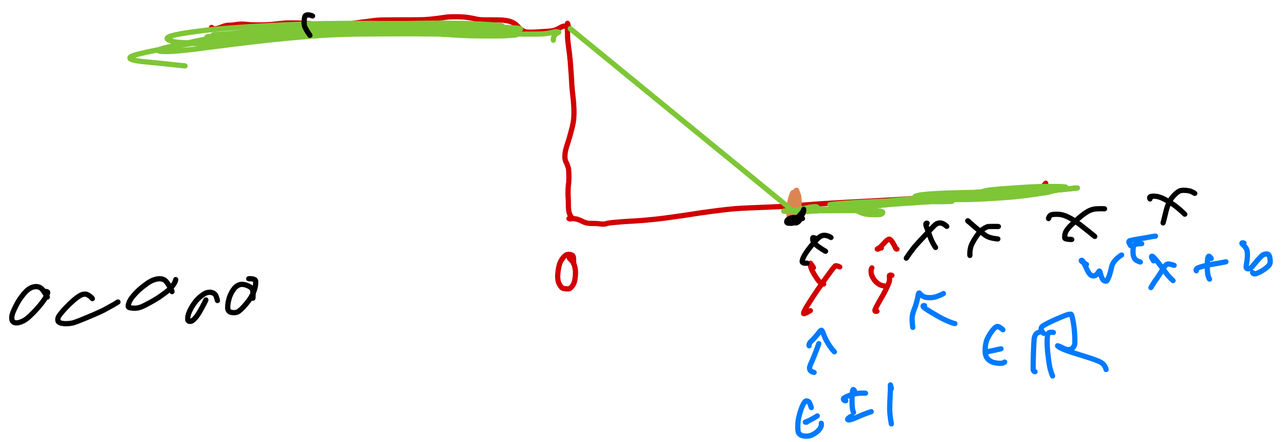


Margin-based bounds

- Would like to analyze 0-1 loss, but it's not Lipschitz, so we can't
  - Not just a technicality!
- We'll analyze with **ramp loss**:

$$\ell^{\text{ramp}}(h,(x,y)) = \psi^{\text{ramp}}(y\,h(x))$$

$$\psi^{\text{ramp}}(z) = \min\left(1, \max\left(0, 1-z\right)\right) = \begin{cases} 1 & \text{if } z < 0 \\ 1-z & \text{if } 0 \le z \le 1 \\ 0 & \text{if } z > 1 \end{cases}$$



Margin-based bounds

- Would like to analyze 0-1 loss, but it's not Lipschitz, so we can't
  - Not just a technicality!
- We'll analyze with **ramp loss**:

$$\ell^{\text{ramp}}(h,(x,y)) = \psi^{\text{ramp}}(y\,h(x))$$

$$\psi^{\text{ramp}}(z) = \min\left(1, \max\left(0, 1-z\right)\right) = \begin{cases} 1 & \text{if } z < 0 \\ 1-z & \text{if } 0 \le z \le 1 \\ 0 & \text{if } z > 1 \end{cases}$$

- ψ^{ramp} is 1-Lipschitz, bounded in $[0,1]$, and $\ell^{0-1} \le \ell^{\text{ramp}}$; but it's *not* convex

Margin-based bounds

- Would like to analyze 0-1 loss, but it's not Lipschitz, so we can't
  - Not just a technicality!
- We'll analyze with **ramp loss**:

$$\ell^{\text{ramp}}(h,(x,y)) = \psi^{\text{ramp}}(y\,h(x))$$

$$\psi^{\text{ramp}}(z) = \min\left(1, \max\left(0, 1-z\right)\right) = \begin{cases} 1 & \text{if } z < 0 \\ 1-z & \text{if } 0 \le z \le 1 \\ 0 & \text{if } z > 1 \end{cases}$$

- ψ^{ramp} is 1-Lipschitz, bounded in $[0,1]$, and $\ell^{0-1} \le \ell^{\text{ramp}}$; but it's *not* convex
- With probability at least $1-\delta$, it holds for all $h \in \mathscr{H}$ that

$$L_{\mathscr{D}}^{\text{ramp}}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{m}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Margin-based bounds

- Would like to analyze 0-1 loss, but it's not Lipschitz, so we can't
  - Not just a technicality!
- We'll analyze with **ramp loss**:

$$\ell^{\text{ramp}}(h,(x,y)) = \psi^{\text{ramp}}(y\,h(x))$$

$$\psi^{\text{ramp}}(z) = \min\left(1, \max\left(0, 1-z\right)\right) = \begin{cases} 1 & \text{if } z < 0 \\ 1-z & \text{if } 0 \le z \le 1 \\ 0 & \text{if } z > 1 \end{cases}$$

- ψ^{ramp} is 1-Lipschitz, bounded in $[0,1]$, and $\ell^{0-1} \le \ell^{\text{ramp}}$; but it's *not* convex
- With probability at least $1-\delta$, it holds for all $h \in \mathscr{H}$ that

$$L_{\mathscr{D}}^{0-1}(h) \le L_{\mathscr{D}}^{\text{ramp}}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \leq L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \leq L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \geq 1) = 1$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \leq L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \geq 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \leq \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \leq R^2$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$
  - By assumption, $\hat{w}_S \in \mathscr{H}$ for hard SVMs

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$
  - By assumption, $\hat{w}_S \in \mathscr{H}$ for hard SVMs
  - Also $L_S^{\text{ramp}}(\hat{w}_S) = 0$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$
  - By assumption, $\hat{w}_S \in \mathscr{H}$ for hard SVMs
  - Also $L_S^{\text{ramp}}(\hat{w}_S) = 0$

$$L_{\mathscr{D}}^{0-1}(\hat{w}_S) \le \frac{2R\|w^*\|}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\rho}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$
  - By assumption, $\hat{w}_S \in \mathscr{H}$ for hard SVMs
  - Also $L_S^{\text{ramp}}(\hat{w}_S) = 0$

$$L_{\mathscr{D}}^{0-1}(\hat{w}_S) \le \frac{2R\|w^*\|}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Got a "slow" $1/\sqrt{n}$ rate

Sample complexity of Hard SVMs

$$L_{\mathscr{D}}^{0-1}(h) \le L_S^{\text{ramp}}(h) + \frac{2}{\gamma}\mathfrak{R}_n\left(\mathscr{H}\right) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Assume (homogenously) realizable with margin 1: $\Pr_{(x,y)\sim\mathscr{D}}(y\langle w^*, x\rangle \ge 1) = 1$
- Let $\mathscr{H} = \left\{x \mapsto w^\top x : \|w\| \le \|w^*\|\right\}$; assume $\mathbb{E}\|x\|^2 \le R^2$
  - $\mathfrak{R}_n(\mathscr{H}) \le R\|w^*\|/\sqrt{n}$
  - By assumption, $\hat{w}_S \in \mathscr{H}$ for hard SVMs
  - Also $L_S^{\text{ramp}}(\hat{w}_S) = 0$

$$L_{\mathscr{D}}^{0-1}(\hat{w}_S) \le \frac{2R\|w^*\|}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Got a "slow" $1/\sqrt{n}$ rate
  - Can get "fast" $1/n$ rate with "local Rademacher complexity"

- $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!

- $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathcal{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$

- $L_{\mathscr{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathscr{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathscr{H}_i$ that

$$L_{\mathscr{D}}^{0-1}(w) \leq L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

- $L_{\mathscr{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathscr{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathscr{H}_i$ that

$$L_{\mathscr{D}}^{0-1}(w) \leq L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

- For any w, take $i_w = \max\{1, \left\lceil \log_2 \frac{\|w\|}{r} \right\rceil\}$: gives $w \in \mathscr{H}_{i_w}$

- $L_{\mathscr{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}} \right) / \sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathscr{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathscr{H}_i$ that

$$L_{\mathscr{D}}^{0-1}(w) \leq L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

(This is a corrected version of Shai+Shai's Theorem 26.14)

- For any w, take $i_w = \max\{1, \left\lceil \log_2 \frac{\|w\|}{r} \right\rceil\}$: gives $w \in \mathscr{H}_{i_w}$

- $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}} \right) / \sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathcal{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathcal{H}_i$ that

$$L_{\mathcal{D}}^{0-1}(w) \leq L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

(This is a corrected version of Shai+Shai's Theorem 26.14)

- For any w, take $i_w = \max\{1, \left\lceil \log_2 \frac{\|w\|}{r} \right\rceil\}$: gives $w \in \mathcal{H}_{i_w}$

$$B_{i_w} \leq \max\{2r, 2\|w\|\}$$

- $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \leq \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathcal{H}_i = \{w : \|w\| \leq B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathcal{H}_i$ that

$$L_{\mathcal{D}}^{0-1}(w) \leq L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

(This is a corrected version of Shai+Shai's Theorem 26.14)

- For any w, take $i_w = \max\{1, \left\lceil \log_2 \frac{\|w\|}{r} \right\rceil\}$: gives $w \in \mathcal{H}_{i_w}$

$$B_{i_w} \leq \max\{2r, 2\|w\|\} \qquad \frac{2}{\delta_{i_w}} = \frac{\pi^2 \left\lceil \max\left\{1, \log_2 \frac{\|w\|}{r}\right\}\right\rceil^2}{3\delta} \leq \frac{14\left(\max\left\{1, \log_2 \frac{\|w\|}{r}\right\}\right)^2}{\delta}$$

- $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \le \left(2R\|w^*\| + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)/\sqrt{n}$ – but we don't know $\|w^*\|$!
- Let $B_i = r2^i$, $\mathcal{H}_i = \{w : \|w\| \le B_i\}$, $\delta_i = \frac{6\delta}{\pi^2 i^2}$ so that $\sum_{i=1}^{\infty} \delta_i = \delta$
- For each i, with prob at least $1 - \delta_i$ have uniformly over $\mathcal{H}_i$ that

$$L_{\mathcal{D}}^{0-1}(w) \le L_S^{\text{ramp}}(w) + \frac{2B_i R}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta_i}}$$

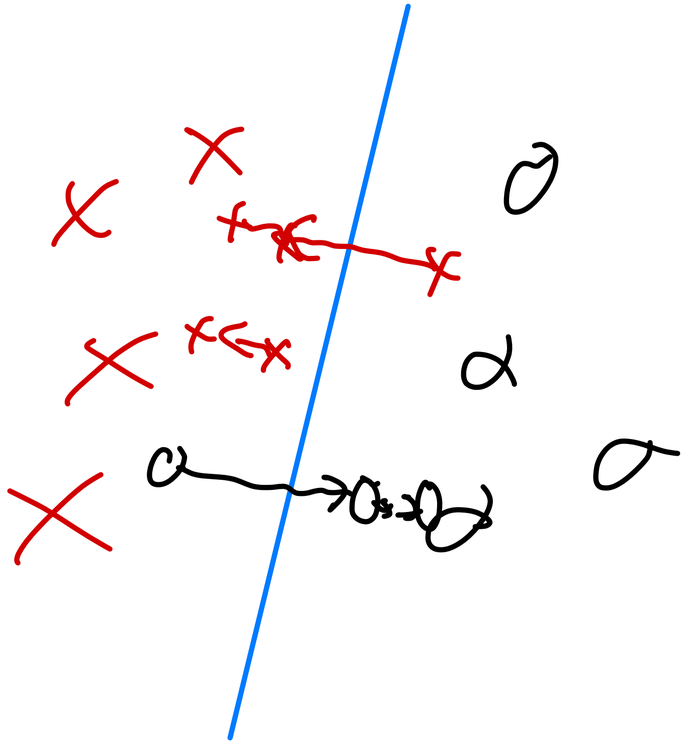
(This is a corrected version of Shai+Shai's Theorem 26.14)

- For any w, take $i_w = \max\{1, \left\lceil \log_2 \frac{\|w\|}{r} \right\rceil\}$: gives $w \in \mathcal{H}_{i_w}$

$$B_{i_w} \le \max\{2r, 2\|w\|\} \qquad \frac{2}{\delta_{i_w}} = \frac{\pi^2 \left\lceil \max\left\{1, \log_2 \frac{\|w\|}{r}\right\}\right\rceil^2}{3\delta} \le \frac{14\left(\max\left\{1, \log_2 \frac{\|w\|}{r}\right\}\right)^2}{\delta}$$

So $L_{\mathcal{D}}^{0-1}(\hat{w}_S) \le \frac{4R\max\{\|\hat{w}_S\|, r\}}{\sqrt{n}} + \sqrt{\frac{1}{n}\max\left\{0, \log\log_2 \frac{\|\hat{w}_S\|}{r}\right\} + \frac{1}{2n}\log\frac{14}{\delta}}$

What if it’s not linearly separable?



What if it's not linearly separable?

- Hard SVM is

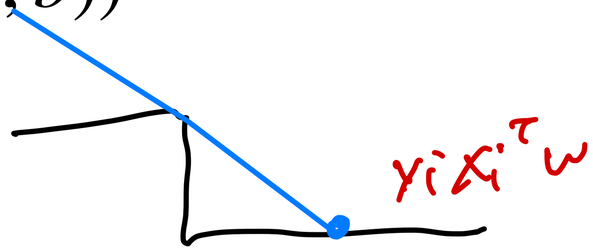
$$\arg\min_{w,b} \|w\|^2 \quad \text{s.t.} \quad \forall i,\ y_i(w^\top x_i + b) \geq 1$$

- **Soft SVM** adds some **slack variables** ξ_i to allow violating the constraints:

$$\arg\min_{w,b,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \forall i,\ y_i(w^\top x_i + b) \geq 1 - \xi_i,\quad \xi_i \geq 0$$

- Equivalent to RLM with $\ell^{\text{hinge}}(h,(x,y)) = \max\{0, 1 - y\,h(x)\}$

$$\arg\min_{w,b} \lambda\|w\|^2 + L_S^{\text{hinge}}((w,b))$$



Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \leq R) = 1$

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \leq R) = 1$
  - Also convex, and we're using an L2 regularizer

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \leq R) = 1$
  - Also convex, and we're using an L2 regularizer
- So, stability analysis from last time tells us

$$\mathbb{E}_S[L^{\text{hinge}}_{\mathscr{D}}(\hat{w}_S)] \leq L^{\text{hinge}}_{\mathscr{D}}(w^*) + \lambda \|w^*\|^2 + \frac{2R^2}{\lambda n}$$

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \leq R) = 1$
  - Also convex, and we're using an L2 regularizer
- So, stability analysis from last time tells us

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \leq \mathbb{E}_S[L_{\mathscr{D}}^{\text{hinge}}(\hat{w}_S)] \leq L_{\mathscr{D}}^{\text{hinge}}(w^*) + \lambda \|w^*\|^2 + \frac{2R^2}{\lambda n}$$

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \leq R) = 1$
  - Also convex, and we're using an L2 regularizer
- So, stability analysis from last time tells us

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \leq \mathbb{E}_S[L_{\mathscr{D}}^{\text{hinge}}(\hat{w}_S)] \leq L_{\mathscr{D}}^{\text{hinge}}(w^*) + \lambda \|w^*\|^2 + \frac{2R^2}{\lambda n}$$

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \leq \inf_{\|w\| \leq B} L_{\mathscr{D}}^{\text{hinge}}(w) + 2RB\sqrt{\frac{2}{n}} \quad \text{if } \lambda = \frac{R}{B}\sqrt{\frac{2}{n}}$$

Generalization bound for Soft-SVM

- Take $b = 0$; then $\ell^{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y\, w^\top x\}$
  - This is $\|x\|$-Lipschitz in w; assume $\Pr(\|x\| \le R) = 1$
  - Also convex, and we're using an L2 regularizer
- So, stability analysis from last time tells us

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \le \mathbb{E}_S[L_{\mathscr{D}}^{\text{hinge}}(\hat{w}_S)] \le L_{\mathscr{D}}^{\text{hinge}}(w^*) + \lambda\|w^*\|^2 + \frac{2R^2}{\lambda n}$$

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \le \inf_{\|w\| \le B} L_{\mathscr{D}}^{\text{hinge}}(w) + 2RB\sqrt{\frac{2}{n}} \quad \text{if } \lambda = \frac{R}{B}\sqrt{\frac{2}{n}}$$

Or ramp loss analysis still works too:

$$\mathscr{L}_{\mathscr{D}}^{0-1}(\hat{w}_S) \le \mathscr{L}_{\mathscr{D}}^{\text{ramp}}(\hat{w}_S) + \frac{2RB}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}} \quad \text{for } \mathscr{H} = \{x \mapsto w^\top x : \|w\| \le B\}$$

$$L_{\mathscr{D}}^{0-1}(\hat{w}_S) \le L_{\mathscr{D}}^{\text{ramp}}(\hat{w}_S) + \frac{4R\max\{\|\hat{w}_S\|, r\}}{\sqrt{n}} + \sqrt{\frac{1}{n}\max\left\{0, \log\log_2\frac{\|\hat{w}_S\|}{r}\right\} + \frac{1}{2n}\log\frac{14}{\delta}}$$

Dimension-free rates

- Ramp loss analysis: if $\mathbb{E}\|x\|^2 \le R^2$,

$$L_{\mathscr{D}}^{0-1}(\hat{w}) \le L_{\mathscr{D}}^{\text{ramp}}(\hat{w}) + \frac{4R\max\{\|\hat{w}_S\|, r\}}{\sqrt{n}} + \sqrt{\frac{1}{n}\max\left\{0, \log\log_2 \frac{\|\hat{w}_S\|}{r}\right\} + \frac{1}{2n}\log\frac{14}{\delta}}$$

- Soft SVM stability analysis: if $\|x\| \le R$ a.s. and $\lambda = \frac{R}{B}\sqrt{\frac{2}{n}}$,

$$\mathbb{E}_S[L_{\mathscr{D}}^{0-1}(\hat{w}_S)] \le \inf_{\|w\|\le B} L_{\mathscr{D}}^{\text{hinge}}(w) + 2RB\sqrt{\frac{2}{n}}$$

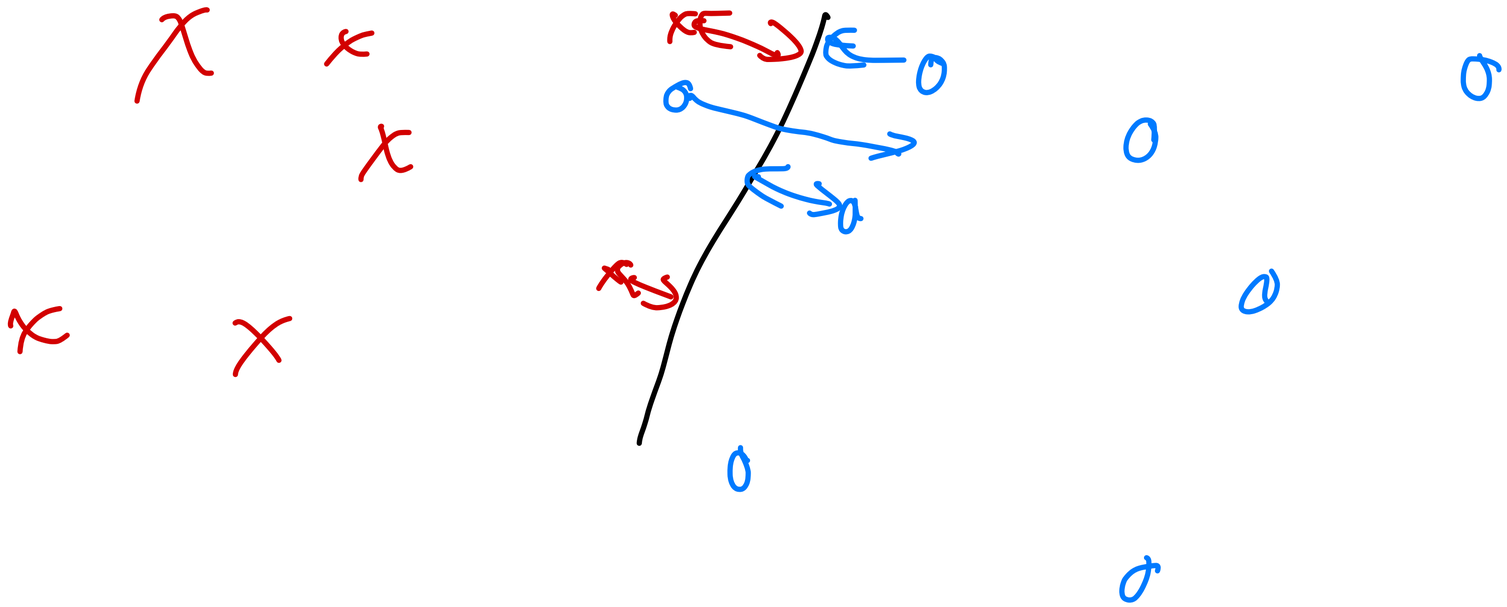
- Neither bound has a d in it! Rate only depends on $\|x\|\|w\| = \frac{\|x\|}{\text{margin}}$
- So we can learn in very high dimensions – even infinite

Wait, but what about $\mathrm{VCdim}(\mathscr{H})$?

- We have $\mathrm{VCdim}(\mathscr{H}) = d$ (in the homogeneous case)
- But these analyses claim we can learn $\mathscr{H}$ in infinite dimensions
  - …because they assumed a margin
  - There exist (even realizable) distributions we can't learn in high-d, but they must have small margin

Why's it called a "support vector machine"?

- At convergence, hinge loss for (hopefully) most training examples will be 0
- The ones where it's not "support" the separating hyperplane



Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1$$

(We didn't actually go through this or the soft SVM version in class; might do next time.)

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$\geq \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

w optimization problem is differentiable + unconstrained

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

$$w = X^\top \operatorname{diag}(y) \alpha$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

$$w = X^\top \operatorname{diag}(y) \alpha \qquad w^\top x = \alpha^\top \operatorname{diag}(y) X x$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

called **weak duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

α_i is zero if $y_i w^\top x_i > 1$

$$w = X^\top \operatorname{diag}(y) \alpha \qquad w^\top x = \alpha^\top \operatorname{diag}(y) X x$$

Soft SVM Duality

$$\min_{w,\xi} \ \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

Only difference from hard SVM is upper bound on $\tilde{\alpha}_i$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

Only difference from hard SVM is upper bound on $\tilde{\alpha}_i$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Can do with b also:
add $\alpha^\top y = 0$ constraint,
set $b = w^\top x_i - y_i$ for any SV

FYI

Karush–Kuhn–Tucker conditions

From Wikipedia, the free encyclopedia

In mathematical optimization, the **Karush–Kuhn–Tucker (KKT) conditions**, also known as the **Kuhn–Tucker conditions**, are first derivative tests (sometimes called first-order necessary conditions) for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.

- Summarize the process of going through Lagrange duality for you
  - Like Lagrange multipliers, but allow inequality constraints
- Make things a lot faster once you're familiar with them
- Related conditions for when strong duality holds
  - Especially important: "Slater's condition"

Practical SVM optimization

- Lot of work in the 90s/00s on efficient SVM dual optimizers
  - Classic implementation: LIBSVM for the dual (wrapped in scikit-learn)
  - Also SVMlight (restrictive license)
  - These days: ThunderSVM
- Primal solvers:
  - LIBLINEAR (wrapped in scikit-learn)
  - SVMperf (restrictive license)
  - Pegasos is among the best, and you already know how to do it
    - It’s just (optionally, projected) stochastic subgradient descent on hinge loss
  - Can still handle kernels this way, just a little less obvious

Logistic regression and margins

- An extremely common classifier: **logistic regression**
  - $\ell^{\text{logistic}}(h, (x, y)) = \log(1 + \exp(-y\, h(x)))$

- On linearly separable data:
  - Limit of low regularization maximizes margin (Rosset, Zhu, Hastie NeurIPS-03)
  - Gradient descent on unregularized problem does too (Soudry et al. JMLR 2018)
    - (related work with other algorithms, other losses, other models…)

- On inseparable data:
  - Gradient descent is biased towards max-margin (Ji and Telgarsky COLT-19)

Summary

- Margin maximization: seems like a reasonable idea
- Hard SVM: exactly maximizes the margin when separable
- Soft SVM: maximizes the margin with some slack
  - equivalent to hinge loss with L2 regularization

- Two analyses:
  - Either case (actually, any linear model): ramp loss with Rademacher
  - Soft SVM: stability analysis
- Rates for both based on the margin, **not** ambient dimension

- Dual form:
  - Shows optimal w can be written as linear combination of training x_i
  - Can be helpful computationally if $n \ll d$
  - Motivates the **kernel trick** – next time!