Boosting with the Logistic Loss is Consistent
Boosting with the Logistic Loss is Consistent

Boring Goal:
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:**
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum,
Boring Goal: Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

Difficulty: No regularization / constraints,
no minimizers / strong convexity,
linearly dependent features / singular Hessian,
infinite dimension / nasty spectrum,
Lipschitz $\implies$ small Hessian on bad errors.
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum, Lipschitz $\implies$ small Hessian on bad errors.

...
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum, Lipschitz $\implies$ small Hessian on bad errors.

... *why bother?*
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum, Lipschitz $\implies$ small Hessian on bad errors.

... *why bother?*

**Aspiration:**
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints,
no minimizers / strong convexity,
linearly dependent features / singular Hessian,
infinite dimension / nasty spectrum,
Lipschitz $\implies$ small Hessian on bad errors.

... *why bother?*

**Aspiration:** Reusable techniques for similar problems.
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum, Lipschitz $\implies$ small Hessian on bad errors.

... why bother?

**Aspiration:** Reusable techniques for similar problems.

**Strategy:**
Boosting with the Logistic Loss is Consistent

**Boring Goal:** Statistical rates for AdaBoost with Logistic and similar strictly convex Lipschitz losses.

**Difficulty:** No regularization / constraints, no minimizers / strong convexity, linearly dependent features / singular Hessian, infinite dimension / nasty spectrum, Lipschitz $\Rightarrow$ small Hessian on bad errors.

... why bother?

**Aspiration:** Reusable techniques for similar problems.

**Strategy:** Identify structure over source distribution via duality; carry it to sample.
Nonseparable case

- When $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,
Nonseparable case

- When $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

\[
\inf \left\{ \text{Logistic loss of } f : f \in \text{span}(\mathcal{H}) \right\}
\]

- When optimal value positive: dual optimum certifies difficulty in every direction.

- Difficulty in every direction $\Rightarrow$ norm constraints.

- Rate $m - c$; $c$ depends on $\mathcal{H}$ and $\mu$. 
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

$$\inf \left\{ \int \ell( -y f(x) ) \, d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}$$
Nonseparable case

- When \( \ell : \mathbb{R} \to \mathbb{R}_+ \) is nondecreasing, \( \beta \)-Lipschitz,

\[
\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\} = \max \left\{ \text{Fermi-Dirac entropy of } p \right\}
\]
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

  $$\inf \left\{ \int \ell(-yf(x)) d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}$$

  $$= \max \left\{ -\int \ell^*(p(x,y)) d\mu(x,y) : \right\}$$

- When optimal value positive: dual optimum certifies difficulty in every direction.

- Difficulty in every direction $\Rightarrow$ norm constraints.

- Rate $m-c$; $c$ depends on $\mathcal{H}$ and $\mu$. 
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

\[
\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\} = \max \left\{ -\int \ell^*(p(x,y))d\mu(x,y) : p \text{ has capped weights}, \right\}
\]
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

$$\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}$$

$$= \max \left\{ -\int \ell^*(p(x,y))d\mu(x,y) : p \in L^1(\mu), p \in [0, \beta] \mu\text{-a.e.,} \right\}$$
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

$$
\inf \left\{ \int \ell( -y f(x) ) d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}
$$

$$
= \max \left\{ - \int \ell^*(p(x,y)) d\mu(x,y) : p \in L^1(\mu), \ p \in [0, \beta] \ \mu\text{-a.e.}, \ p \text{ decorrelates } \mathcal{H} \text{ from } \mu \right\}
$$
Nonseparable case

When \( \ell : \mathbb{R} \to \mathbb{R}_+ \) is nondecreasing, \( \beta \)-Lipschitz,

\[
\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}
\]

\[
= \max \left\{ -\int \ell^*(p(x,y))d\mu(x,y) : p \in L^1(\mu), p \in [0, \beta] \, \mu\text{-a.e.}, \right. \\
\quad \left. \forall f \in \text{span}(\mathcal{H}) \cdot \int yf(x)p(x,y)d\mu(x,y) = 0 \right\}.
\]
Nonseparable case

- When $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

$$\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}$$

$$= \max \left\{ -\int \ell^*(p(x,y))d\mu(x,y) : p \in L^1(\mu), p \in [0, \beta] \text{ } \mu\text{-a.e.}, \forall f \in \text{span}(\mathcal{H}) \cdot \int yf(x)p(x,y)d\mu(x,y) = 0 \right\}.$$  

- When optimal value positive:

  Difficulty in every direction $\Rightarrow$ norm constraints.

  Rate $m - c$; $c$ depends on $\mathcal{H}$ and $\mu$. 

  $\forall f \in \text{span}(\mathcal{H}) \cdot \int yf(x)p(x,y)d\mu(x,y) = 0$. 

- When optimal value positive:
Nonseparable case

- When \( \ell : \mathbb{R} \to \mathbb{R}_+ \) is nondecreasing, \( \beta \)-Lipschitz,

\[
\inf \left\{ \int \ell( -y f(x) ) d\mu(x, y) : f \in \text{span}(\mathcal{H}) \right\}
\]

\[
= \max \left\{ -\int \ell^*(p(x, y)) d\mu(x, y) : p \in L^1(\mu), p \in [0, \beta] \ \mu\text{-a.e.}, \right. \\
\left. \forall f \in \text{span}(\mathcal{H}) \cdot \int y f(x)p(x, y) d\mu(x, y) = 0 \right\}.
\]

- When optimal value positive: dual optimum certifies difficulty in every direction.
Nonseparable case

- When $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

\[
\inf \left\{ \int \ell(-y f(x)) d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\} = \max \left\{ -\int \ell^*(p(x,y)) d\mu(x,y) : p \in L^1(\mu), p \in [0, \beta] \mu\text{-a.e.}, \right. \\
\left. \forall f \in \text{span}(\mathcal{H}) \cdot \int y f(x) p(x,y) d\mu(x,y) = 0 \right\}.
\]

- When optimal value positive: dual optimum certifies difficulty in every direction.
- Difficulty in every direction $\implies$ norm constraints.
Nonseparable case

- When $\ell: \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, $\beta$-Lipschitz,

$$\inf \left\{ \int \ell(-yf(x))d\mu(x,y) : f \in \text{span}(\mathcal{H}) \right\}$$

$$= \max \left\{ -\int \ell^*(p(x,y))d\mu(x,y) : p \in L^1(\mu), p \in [0, \beta] \mu\text{-a.e.}, \right. \left. \forall f \in \text{span}(\mathcal{H}) \cdot \int yf(x)p(x,y)d\mu(x,y) = 0 \right\}.$$

- When optimal value positive:
  dual optimum certifies difficulty in every direction.
- Difficulty in every direction $\implies$ norm constraints.
- Rate $m^{-c}$; $c$ depends on $\mathcal{H}$ and $\mu$.
Separable case

- What if optimal value zero?
Separable case

- What if optimal value zero?
- Weak learning rate:

\[ \gamma = \]
Separable case

- What if optimal value zero?
- Weak learning rate:

\[ \gamma = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H}) \atop \|f\|_\star = 1} \int y f(x)p(x, y)d\mu(x, y) \right\} \]

...!
Separable case

- What if optimal value zero?
- Weak learning rate:

\[
\gamma = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \int yf(x)p(x, y)d\mu(x, y) \middle| \|f\|_* = 1 \right\}
\]

: \( p \in L^1(\mu), \|p\|_1 = 1, \)
Separable case

- What if optimal value zero?
- Weak learning rate:

\[
\gamma = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \right. \\
\left. \int yf(x)p(x, y)d\mu(x, y) \right. \\
\left. : f \in \text{span}(\mathcal{H}), \|f\|_\star = 1 \right. \\
: p \in L^1(\mu), \|p\|_1 = 1, p \in [0, \infty) \mu\text{-a.e.} \right\}.
\]
Separable case

- What if optimal value zero?
- Weak learning rate:

\[
\gamma = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \int yf(x)p(x, y)d\mu(x, y) \middle| \|f\|_* = 1 \right\}
\]

: \( p \in L^1(\mu), \|p\|_1 = 1, p \in [0, \infty] \mu\text{-a.e.} \)
Separable case

- What if optimal value zero?
- Weak learning rate (adapted to Lipschitz losses):

$$\gamma_{\epsilon} = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H}), \|f\|_* = 1} \int y f(x) p(x, y) d\mu(x, y) \right\}.$$

: $p \in L^1(\mu), \|p\|_1 = 1, p \in [0, 1/\epsilon]$ $\mu$-a.e.
Separable case

- What if optimal value zero?
- Weak learning rate (adapted to Lipschitz losses):

\[
\gamma_\epsilon = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \int y f(x) p(x, y) d\mu(x, y) \middle| \|f\|_* = 1 \right\}
\]

\[
: p \in L^1(\mu), \|p\|_1 = 1, p \in [0, 1/\epsilon] \mu\text{-a.e.}
\]

- (Hi Manfred, Rocco, Satyen, Shai, ...)
Separable case

- What if optimal value zero?
- Weak learning rate (adapted to Lipschitz losses):

\[ \gamma_\epsilon = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \int y f(x) p(x, y) d\mu(x, y) \right\} \]

\[
\sup_{\|f\|_* = 1} \int y f(x) p(x, y) d\mu(x, y)
\]

: \( p \in L^1(\mu), \|p\|_1 = 1, p \in [0, 1/\epsilon] \) \( \mu \)-a.e.

- (Hi Manfred, Rocco, Satyen, Shai, ...)
- Earlier optimal value zero \( \iff \gamma_\epsilon > 0 \) for \( \epsilon > 0 \)...!
Separable case

- What if optimal value zero?
- Weak learning rate *(adapted to Lipschitz losses)*:

\[
\gamma_\epsilon = \inf \left\{ \sup_{f \in \text{span}(\mathcal{H})} \int y f(x) p(x, y) d\mu(x, y) \right. \\
\left. \quad \|f\|_\ast = 1 \right\}
\]

\[ : p \in L^1(\mu), \|p\|_1 = 1, p \in [0, 1/\epsilon] \mu\text{-a.e.} \}

- (Hi Manfred, Rocco, Satyen, Shai, ...)
- Earlier optimal value zero $\iff \gamma_\epsilon > 0$ for $\epsilon > 0$...!
- $\gamma_\epsilon$ lower bounds progress; rate $\mathcal{O}(m^{-1/3})$. 