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,

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

Nonseparable case

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

$$\inf \left\{ \text{Logistic loss of } f \quad : f \in \operatorname{span}(\mathcal{H}) \right\}$$

Nonseparable case

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

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

Nonseparable case

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

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

$$= \max \left\{ \text{Fermi-Dirac entropy of } p \right.$$

Nonseparable case

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

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

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

$$:$$

Nonseparable case

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

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

$$= \max \Bigg\{ -\int \ell^*(p(x,y)) d\mu(x,y)$$

$$: p \text{ has capped weights,}$$

Nonseparable case

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

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

Nonseparable case

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

$$
\begin{aligned}
&\inf\left\{ \int \ell(\ -yf(x)\ )d\mu(x,y) : f \in \mathrm{span}(\mathcal{H}) \right\} \\
&= \max\left\{ -\int \ell^*(p(x,y))d\mu(x,y) \right. \\
&\qquad\qquad : p \in L^1(\mu),\ p \in [0,\beta]\ \mu\text{-a.e.}, \\
&\qquad\qquad\qquad \left. p \text{ decorrelates } \mathcal{H} \text{ from } \mu \right\}
\end{aligned}
$$

Nonseparable case

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

$$
\begin{aligned}
&\inf \left\{ \int \ell(\ -yf(x)\ ) d\mu(x,y) : f \in \operatorname{span}(\mathcal{H}) \right\} \\
&= \max \Bigg\{ -\int \ell^*(p(x,y)) d\mu(x,y) \\
&\qquad\qquad : p \in L^1(\mu),\ p \in [0,\beta]\ \mu\text{-a.e.}, \\
&\qquad\qquad\quad \forall f \in \operatorname{span}(\mathcal{H})\,.\int y f(x) p(x,y) d\mu(x,y) = 0 \Bigg\}.
\end{aligned}
$$

Nonseparable case

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

$$
\begin{aligned}
&\inf \left\{ \int \ell(\ -yf(x)\ ) d\mu(x,y) : f \in \operatorname{span}(\mathcal{H}) \right\} \\
&= \max \Bigg\{ -\int \ell^*(p(x,y)) d\mu(x,y) \\
&\qquad\qquad : p \in L^1(\mu),\ p \in [0,\beta]\ \mu\text{-a.e.}, \\
&\qquad\qquad\quad \forall f \in \operatorname{span}(\mathcal{H}) \centerdot \int y f(x) p(x,y) d\mu(x,y) = 0 \Bigg\}.
\end{aligned}
$$

- When optimal value positive:

Nonseparable case

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

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

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

Nonseparable case

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

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

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

Nonseparable case

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

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

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

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:

$$\gamma = \inf \left\{ \sup_{\substack{f \in \mathrm{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y) \right.$$

$$\vdots$$

Separable case

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

$$\gamma = \inf \left\{ \sup_{\substack{f \in \mathrm{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y) \right.$$

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

Separable case

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

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

Separable case

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

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

Separable case

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

$$\gamma_\epsilon = \inf \Bigg\{ \sup_{\substack{f \in \text{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y)$$
$$: p \in L^1(\mu),\ \|p\|_1 = 1,\ p \in [0, 1/\epsilon]\ \mu\text{-a.e.} \Bigg\}.$$

Separable case

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

$$\gamma_\epsilon = \inf \left\{ \sup_{\substack{f \in \operatorname{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y) \right.$$
$$\left. : p \in L^1(\mu),\ \|p\|_1 = 1,\ p \in [0, 1/\epsilon]\ \mu\text{-a.e.} \right\}.$$

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

Separable case

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

$$\gamma_\epsilon = \inf \left\{ \sup_{\substack{f \in \mathrm{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y) \right.$$
$$\left. : p \in L^1(\mu),\ \|p\|_1 = 1,\ p \in [0, 1/\epsilon]\ \mu\text{-a.e.} \right\}.$$

- (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):

$$\gamma_\epsilon = \inf \left\{ \sup_{\substack{f \in \operatorname{span}(\mathcal{H}) \\ \|f\|_\star = 1}} \int y f(x) p(x,y) d\mu(x,y) \right.$$
$$\left. : p \in L^1(\mu),\ \|p\|_1 = 1,\ p \in [0, 1/\epsilon]\ \mu\text{-a.e.} \right\}.$$

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