Fundamental Perspectives

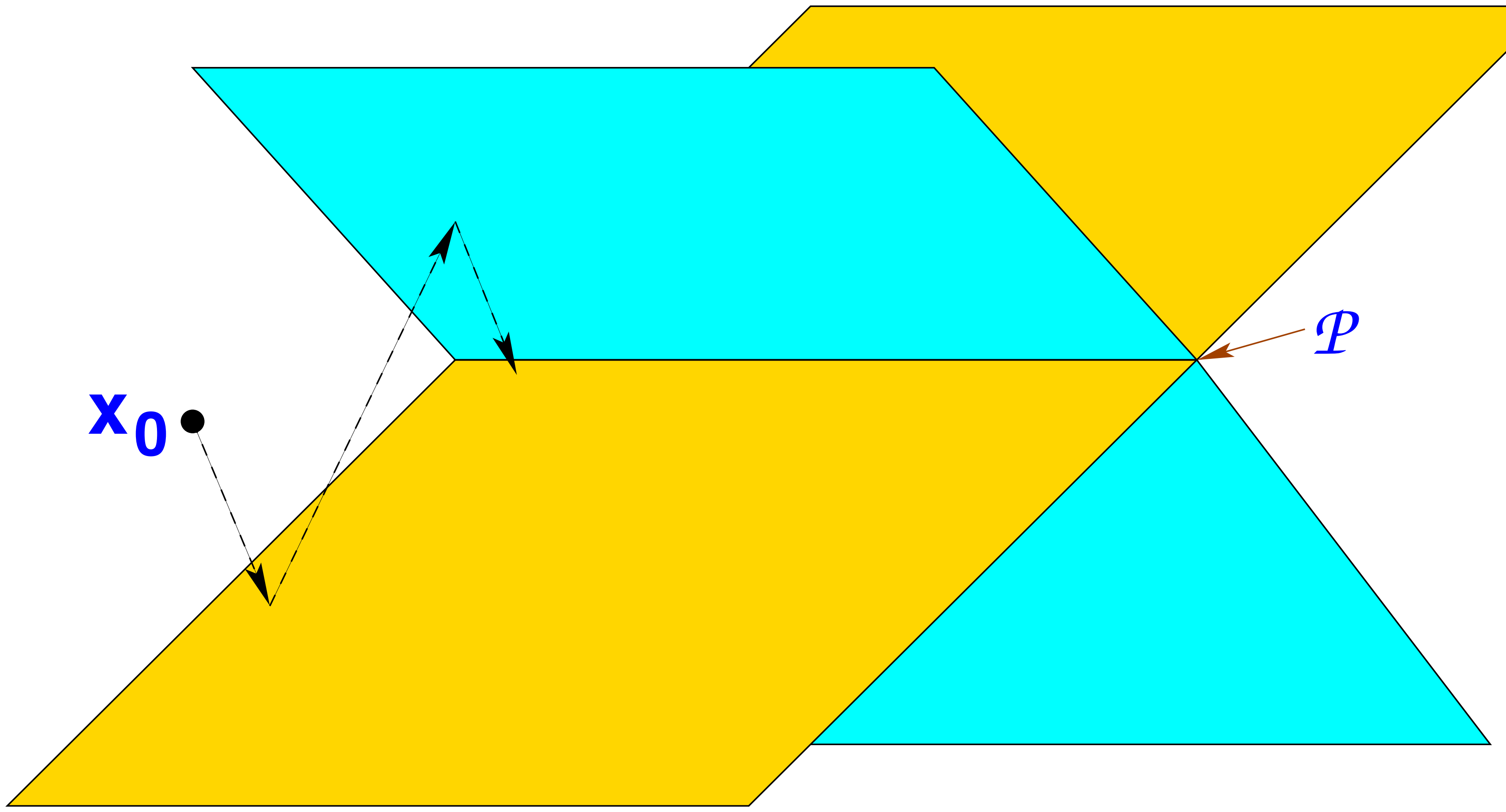
- game theory
- loss minimization
- an information-geometric view

A Dual Information-Geometric Perspective

- loss minimization focuses on function computed by AdaBoost (i.e., weights on weak classifiers)
- dual view: instead focus on distributions D_t (i.e., weights on examples)
- dual perspective combines geometry and information theory
- exposes underlying mathematical structure
- basis for proving convergence

An Iterative-Projection Algorithm

- say want to find point closest to $\mathbf{x}_0$ in set
  $\mathcal{P} = \{$ intersection of N hyperplanes $\}$
- algorithm: [Bregman; Censor & Zenios]
  - start at $\mathbf{x}_0$
  - repeat: pick a hyperplane and project onto it



- if $\mathcal{P} \neq \emptyset$, under general conditions, will converge correctly

AdaBoost is an Iterative-Projection Algorithm

[Kivinen & Warmuth]

- points = distributions D_t over training examples
- distance = relative entropy:

$$\mathrm{RE}\left(P \parallel Q\right) = \sum_i P(i) \ln\left(\frac{P(i)}{Q(i)}\right)$$

- reference point $\mathbf{x}_0 =$ uniform distribution
- hyperplanes defined by all possible weak classifiers g_j:

$$\sum_i D(i) y_i g_j(x_i) = 0 \Leftrightarrow \Pr_{i \sim D}\left[g_j(x_i) \neq y_i\right] = \tfrac{1}{2}$$

  - intuition: looking for "hardest" distribution

AdaBoost as Iterative Projection (cont.)

- algorithm:
  - start at D_1 = uniform
  - for $t = 1, 2, \ldots$:
    - pick hyperplane/weak classifier $h_t \leftrightarrow g_j$
    - D_{t+1} = (entropy) projection of D_t onto hyperplane

$$= \arg \min_{D : \sum_i D(i) y_i g_j(x_i) = 0} \mathrm{RE}(D \parallel D_t)$$

- claim: equivalent to AdaBoost
- further: choosing h_t with minimum error $\equiv$ choosing farthest hyperplane

Boosting as Maximum Entropy

- corresponding optimization problem:

$$\min_{D \in \mathcal{P}} \mathrm{RE}\left(D \;\|\; \text{uniform}\right) \leftrightarrow \max_{D \in \mathcal{P}} \text{entropy}(D)$$

- where

$$\begin{aligned} \mathcal{P} &= \text{feasible set} \\ &= \left\{ D : \sum_i D(i) y_i g_j(x_i) = 0 \;\; \forall j \right\} \end{aligned}$$

- $\mathcal{P} \neq \emptyset \Leftrightarrow$ weak learning assumption does not hold
  - in this case, $D_t \rightarrow$ (unique) solution
- if weak learning assumption does hold then
  - $\mathcal{P} = \emptyset$
  - D_t can never converge
  - dynamics are fascinating but unclear in this case

Unifying the Two Cases

[Schapire Collins & Singer]

- two distinct cases:
  - weak learning assumption holds
    - $\mathcal{P} = \emptyset$
    - dynamics unclear
  - weak learning assumption does not hold
    - $\mathcal{P} \neq \emptyset$
    - can prove convergence of D_t's
- to unify: work instead with unnormalized versions of D_t's
  - standard AdaBoost: $D_{t+1}(i) = \dfrac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{\text{normalization}}$
  - instead:

$$d_{t+1}(i) = d_t(i) \exp(-\alpha_t y_i h_t(x_i))$$

$$D_{t+1}(i) = \frac{d_{t+1}(i)}{\text{normalization}}$$

  - algorithm is unchanged

Reformulating AdaBoost as Iterative Projection

- points = nonnegative vectors $\mathbf{d}_t$
- distance = unnormalized relative entropy:

$$\mathrm{RE}\left(\mathbf{p} \parallel \mathbf{q}\right) = \sum_i \left[p(i) \ln\left(\frac{p(i)}{q(i)}\right) + q(i) - p(i)\right]$$

- reference point $\mathbf{x}_0 = \mathbf{1}$ (all 1's vector)
- hyperplanes defined by weak classifiers g_j:

$$\sum_i d(i) y_i g_j(x_i) = 0$$

- resulting iterative-projection algorithm is again equivalent to AdaBoost

Reformulated Optimization Problem

- optimization problem:

$$\min_{\mathbf{d}\in\mathcal{P}} \mathrm{RE}\left(\mathbf{d} \;\|\; \mathbf{1}\right)$$

- where

$$\mathcal{P} = \left\{ \mathbf{d} : \sum_i d(i) y_i g_j(x_i) = 0 \;\; \forall j \right\}$$

- note: feasible set $\mathcal{P}$ never empty (since $\mathbf{0} \in \mathcal{P}$)

Exponential Loss as Entropy Optimization

- all vectors $\mathbf{d}_t$ created by AdaBoost have form:

$$d(i) = \exp\left(-y_i \sum_j \lambda_j g_j(x_i)\right)$$

- let $\mathcal{Q} = \{$ all vectors $\mathbf{d}$ of this form $\}$
- can rewrite exponential loss:

$$\begin{aligned}
\inf_{\boldsymbol{\lambda}} \sum_i \exp\left(-y_i \sum_j \lambda_j g_j(x_i)\right) &= \inf_{\mathbf{d}\in\mathcal{Q}} \sum_i d(i) \\
&= \min_{\mathbf{d}\in\overline{\mathcal{Q}}} \sum_i d(i) \\
&= \min_{\mathbf{d}\in\overline{\mathcal{Q}}} \mathrm{RE}\left(\mathbf{0} \;\|\; \mathbf{d}\right)
\end{aligned}$$

- $\overline{\mathcal{Q}}$ = closure of $\mathcal{Q}$

Duality

[Della Pietra, Della Pietra & Lafferty]

- presented two optimization problems:
  - $\min_{\mathbf{d}\in\mathcal{P}} \mathrm{RE}\left(\mathbf{d} \;\|\; \mathbf{1}\right)$
  - $\min_{\mathbf{d}\in\overline{\mathcal{Q}}} \mathrm{RE}\left(\mathbf{0} \;\|\; \mathbf{d}\right)$
- which is AdaBoost solving? Both!
- problems have same solution
- moreover: solution given by unique point in $\mathcal{P} \cap \overline{\mathcal{Q}}$
- problems are convex duals of each other

Convergence of AdaBoost

- can use to prove AdaBoost converges to common solution of both problems:
  - can argue that $\mathbf{d}^* = \lim \mathbf{d}_t$ is in $\mathcal{P}$
  - vectors $\mathbf{d}_t$ are in $\mathcal{Q}$ always $\Rightarrow \mathbf{d}^* \in \overline{\mathcal{Q}}$

  $\therefore$ $\mathbf{d}^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$

  $\therefore$ $\mathbf{d}^*$ solves both optimization problems
- so:
  - AdaBoost minimizes exponential loss
  - exactly characterizes limit of unnormalized "distributions"
  - likewise for normalized distributions when weak learning assumption does not hold
- also, provides additional link to logistic regression
  - only need slight change in optimization problem

[Schapire, Collins, Singer; Lebannon & Lafferty]

Conclusions

- from different perspectives, AdaBoost can be interpreted as:
    - a method for boosting the accuracy of a weak learner
    - a procedure for maximizing margins
    - an algorithm for playing repeated games
    - a numerical method for minimizing exponential loss
    - an iterative-projection algorithm based on an information-theoretic geometry
- none is entirely satisfactory by itself, but each useful in its own way
- taken together, create rich theoretical understanding
    - connect boosting to other learning problems and techniques
    - provide foundation for versatile set of methods with many extensions, variations and applications

References

Coming soon:

- Robert E. Schapire and Yoav Freund.
  Boosting: Foundations and Algorithms.
  MIT Press, 2012.

Survey articles:

- Ron Meir and Gunnar Rätsch.
  An Introduction to Boosting and Leveraging.
  In *Advanced Lectures on Machine Learning (LNAI2600)*, 2003.
  http://www.boosting.org/papers/MeiRae03.pdf
- Robert E. Schapire.
  The boosting approach to machine learning: An overview.
  In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.
  http://www.cs.princeton.edu/~schapire/boost.html