Learning Bounds for Importance Weighting

Corinna Cortes
Google Research
corinna@google.com

Yishay Mansour
Tel-Aviv University
mansour@tau.ac.il

Mehryar Mohri
Courant & Google
mohri@cims.nyu.edu

Special thanks to Ameet Talwalkar (UC Berkeley)

Motivation

- Importance weighting used in variety of contexts:
  - sample bias correction (e.g., Dudík et al., 2006; Zadrozny et al., 2003; Huang et al., 2006; Sugiyama et al., 2008).
  - domain adaptation.
  - active learning (Beygelzimer et al., 2009).
  - analysis of boosting (Dasgupta and Long, 2003).
- Guarantees?
  - when is importance weighting successful?
  - are there better reweighting techniques than the straightforward standard approach?

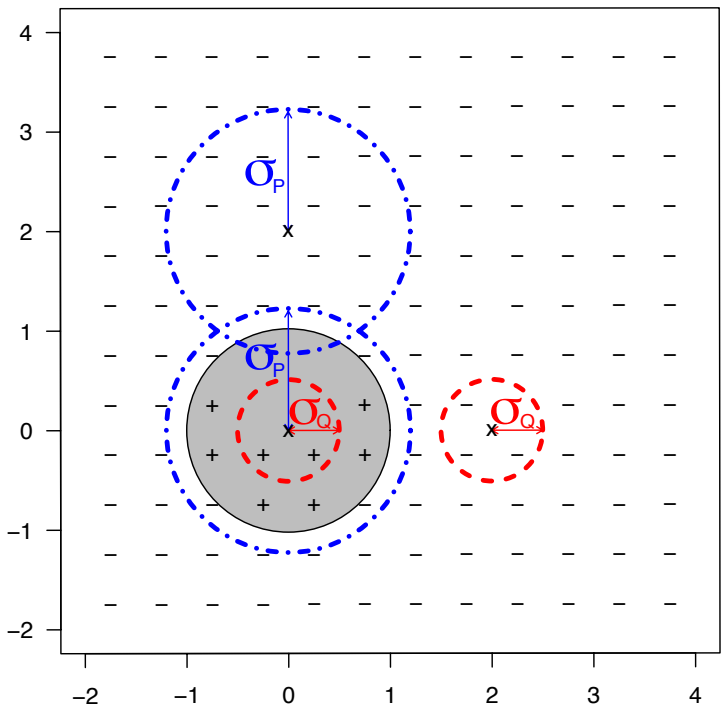
Setting

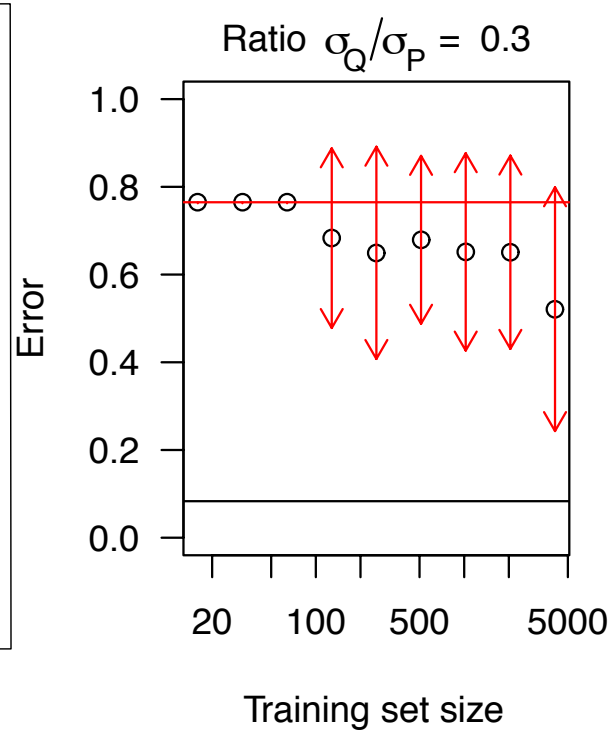
- Input space X, output space Y.
- Loss function $L : Y \times Y \to [0, 1]$.
- Source distribution Q.
- Target distribution P.
- Training sample S of size m drawn according to Q.
- Hypothesis set H.
- Fixed target labeling function $f : X \to Y$.
- Notation: for any $x \in X$ and $h \in H$, $L_h(x) = L(h(x), f(x))$.

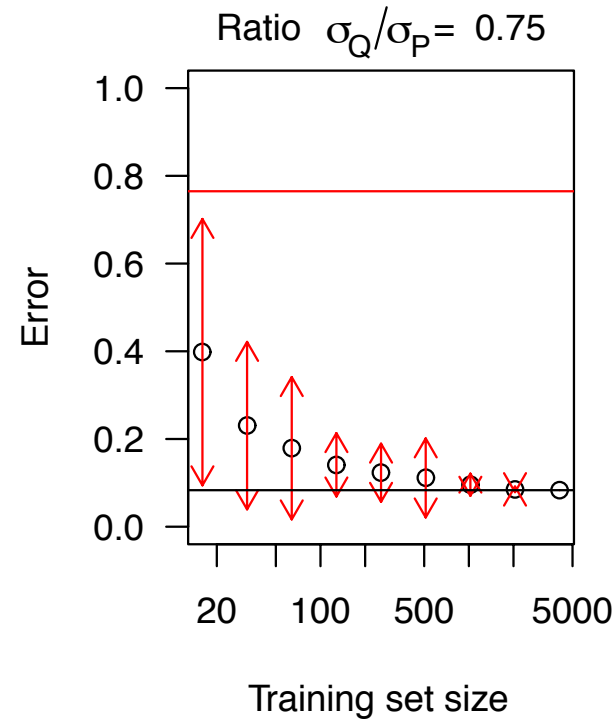
Importance Weighting

- Emphasize loss of training point x by $w(x)$:
  - empirical loss: $\widehat{R}_w(h) = \frac{1}{m} \sum_{i=1}^{m} w(x_i)\, L_h(x_i)$.
  - generalization loss: $R(h) = \mathop{\mathrm{E}}_{x \sim P}[L_h(x)]$.
- Weight assumed known: $w(x) = P(x)/Q(x)$.
  - In practice, estimated weight $\widehat{w}(x)$: effect of this error specifically analyzed by (Cortes et al., 2008).
- Different scenarios: importance weighting/sampling.
  - imp. weighting: finite sample of size $m \sim$ some Q.
  - imp. sampling: unlimited sampling, can choose Q.

Does Importance Weighting Work?







- Hypothesis class: hyperplanes tangent to the unit circle.
- Best hypothesis chosen by empirical risk minimization.

Outline

- Preliminaries.
- Learning bounds for bounded importance weights.
- Learning bounds for unbounded importance weights (the most common case).
- Alternative reweighting techniques.

Rényi Divergences

(Rényi, 1960)

- Definition: for $\alpha \geq 0$,

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log_2 \sum_x P(x) \left[\frac{P(x)}{Q(x)}\right]^{\alpha-1}.$$

  - $D_\alpha \geq 0$ for all $\alpha \geq 0$.
  - D_1 coincides with the relative entropy (KL div.).
  - D_α non-decreasing function of α.

- Notation:

$$d_\alpha(P\|Q) = 2^{D_\alpha(P\|Q)} = \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)}\right]^{\frac{1}{\alpha-1}}.$$

Properties

- Properties of w :

$$\mathrm{E}[w] = 1 \quad \mathrm{E}[w^2] = d_2(P\|Q) \quad \sigma^2(w) = d_2(P\|Q) - 1.$$

- Properties of wL_h:

$$\mathop{\mathrm{E}}_{Q}[\widehat{R}_w(h)] = \mathop{\mathrm{E}}_{x\sim Q}[w(x)L_h(x)] = \sum_x \frac{P(x)}{Q(x)} L_h(x)\, Q(x) = \sum_x P(x) L_h(x) = R(h).$$

$$\mathop{\mathrm{E}}_{x\sim Q}[w^2(x)\, L_h^2(x)] = \sum_x Q(x) \left[\frac{P(x)}{Q(x)}\right]^2 L_h^2(x) = \sum_x P(x)^{\frac{1}{\alpha}} \left[\frac{P(x)}{Q(x)}\right] P(x)^{\frac{\alpha-1}{\alpha}}\, L_h^2(x)$$

$$\text{(Hölder's inequality)} \leq \left[\sum_x P(x) \left[\frac{P(x)}{Q(x)}\right]^{\alpha}\right]^{\frac{1}{\alpha}} \left[\sum_x P(x)\, L_h^{\frac{2\alpha}{\alpha-1}}(x)\right]^{\frac{\alpha-1}{\alpha}}$$

$$= d_{\alpha+1}(P\|Q) \left[\sum_x P(x)\, L_h(x) L_h^{\frac{\alpha+1}{\alpha-1}}(x)\right]^{\frac{\alpha-1}{\alpha}} \leq d_{\alpha+1}(P\|Q)\, R(h)^{1-\frac{1}{\alpha}}.$$

Learning Bound - Bounded Case

- Assumption (bounded case):

$$M = \sup_x w(x) = \sup_x \frac{P(x)}{Q(x)} = d_\infty(P\|Q) < +\infty.$$

- Theorem: let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}_w(h) + \frac{2M(\log|H| + \log\frac{1}{\delta})}{3m} + \sqrt{\frac{2d_2(P\|Q)(\log|H| + \log\frac{1}{\delta})}{m}}.$$

  - similar result for infinite hypothesis sets.
  - note the role of Rényi divergence.

Lower Bound - Bounded Case

- Theorem: assume that $M < \infty$ and $\sigma^2(w)/M^2 \geq 1/m$. Assume that H contains a hypothesis h_0 such that $L_{h_0}(x) = 1$ for all x. Then, there exists an absolute constant $c = 2/41^2$ such that

$$\Pr\left[\sup_{h \in H} \left|R(h) - \widehat{R}_w(h)\right| \geq \sqrt{\frac{d_2(P\|Q) - 1}{4m}}\right] \geq c > 0.$$

  - result based on proof of general lower bound theorem for maximal variance.

Unbounded Case

- Assumption $d_\infty(P\|Q) < \infty$ does not hold, even in some natural cases.
- Examples: Gaussian distributions.

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma_P} \exp\left[-\frac{(x-\mu)^2}{2\sigma_P^2}\right] \quad Q(x) = \frac{1}{\sqrt{2\pi}\sigma_Q} \exp\left[-\frac{(x-\mu')^2}{2\sigma_Q^2}\right].$$

  - even for $\sigma_P = \sigma_Q$ and $\mu \neq \mu'$, $d_\infty(P\|Q) = +\infty$.
  - but, for $\sigma_Q > \frac{\sqrt{2}}{2}\sigma_P$ (e.g., example on the right, slide 2), $d_2(P\|Q) < +\infty$, thus the second-moment of w is bounded.

Learning Bound - Unbounded Case

- Theorem: let H such that $\mathrm{Pdim}(\{L_h(x) : h \in H\}) = p$ is finite. Assume that $d_2(P\|Q) < +\infty$ and $w(x) \neq 0$ for all x. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $h \in H$,

$$R(h) \leq \widehat{R}_w(h) + 2^{5/4}\sqrt{d_2(P\|Q)}\sqrt[8]{\frac{p\log\frac{2me}{p} + \log\frac{4}{\delta}}{m}}.$$

  - holds even for unbounded weights.
  - based on new proof of general learning bound theorem for unbounded losses with bounded second-moment (Vapnik's proof is incorrect!).

Alternative Weighting Techniques

- Arbitrary $u : X \to \mathbb{R}$ with $u > 0$ and for any $h \in H$,

$$\widehat{R}_u(h) = \frac{1}{m} \sum_{i=1}^{m} u(x_i) L_h(x_i).$$

- Theorem: let H such that $\mathrm{Pdim}(\{L_h(x) : h \in H\}) = p$ is finite. Assume that $0 < \mathrm{E}_Q[u^2(x)] < +\infty$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$,

$$|R(h) - \widehat{R}_u(h)| \leq \left| \mathrm{E}_Q \left[[w(x) - u(x)] L_h(x) \right] \right| +$$
$$2^{5/4} \max\left(\sqrt{\mathrm{E}_Q[u^2(x) L_h^2(x)]}, \sqrt{\mathrm{E}_{\widehat{Q}}[u^2(x) L_h^2(x)]} \right) \sqrt[\frac{3}{8}]{\frac{p \log \frac{2me}{p} + \log \frac{4}{\delta}}{m}}.$$

Alternative Weighting Techniques

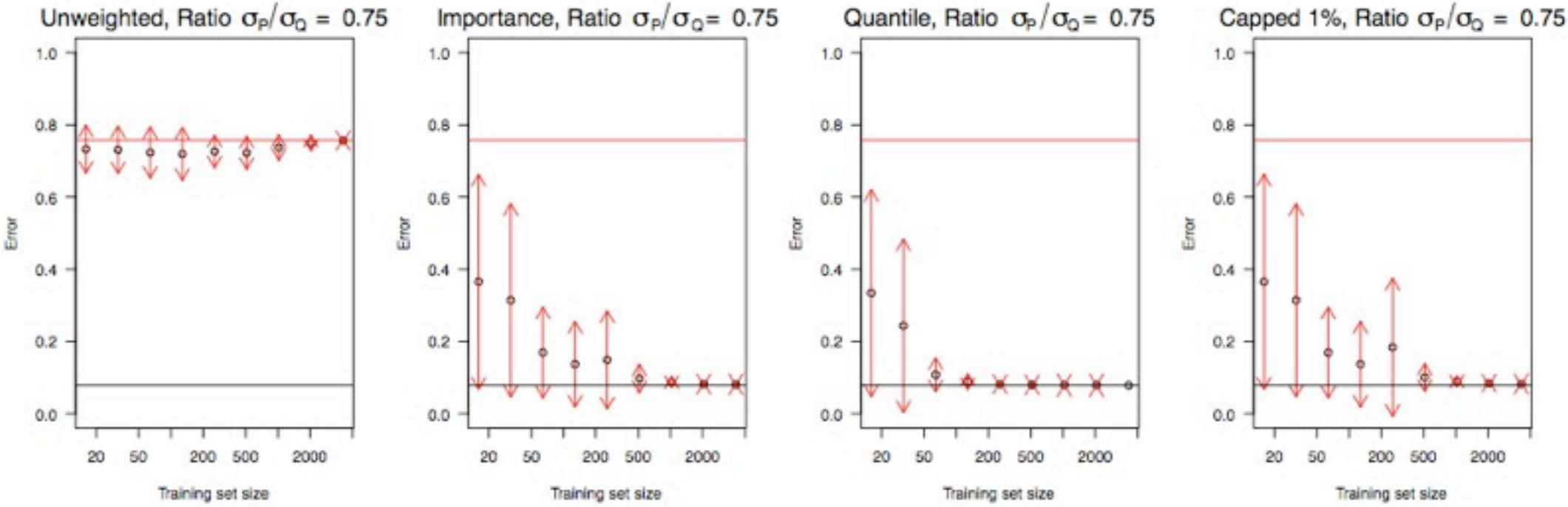
- Trade-off between bias term $|\mathrm{E}_Q[(w(x)-u(x))L_h(x)]|$ and second moment $\sqrt{\max\left(\mathrm{E}_Q[u^2(x)L_h^2(x)], \mathrm{E}_{\widehat{Q}}[u^2(x)L_h^2(x)]\right)}$.
- Using upper bound independent of H leads to the optimization problem

$$\min_{u\in U} \; \mathop{\mathrm{E}}_{Q}\left[|w(x)-u(x)|\right] + \gamma\sqrt{\mathop{\mathrm{E}}_{Q}[u^2]},$$

with $\gamma > 0$ a trade-off parameter.

Alternative Reweighting - Example

(1) (2) (3) (4)



The variance is reduced in (3) by replacing w with the average weight per quantile.

Conclusion and Open Questions

- Learning guarantees for importance weighting, including unbounded case (most common).
- Analysis of cases where importance weighting can succeed.
- Critical role of Rényi divergence of the distributions.
- Preliminary exploration of other reweighting techniques.
- Estimation of Rényi divergence from finite samples.