Passive Learning with Target Risk

Mehrdad Mahdavi       Rong Jin

Department of Computer Science
Michigan State University

COLT 2013
Statistical Learning Theory

Setting:

- The instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- The unknown probability distribution $\mathcal{D}$
- The hypotheses class $\mathcal{H}$
- The loss function $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$

Sample Complexity:

$n(\varepsilon, \delta) : (0; 1) \times (0; 1) \rightarrow \mathbb{N}$, the number of examples required to achieve $\varepsilon$ accuracy with probability at least $1 - \delta$. 
Setting:

- The instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- The unknown probability distribution $\mathcal{D}$
- The hypotheses class $\mathcal{H}$
- The loss function $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$

Given: $\mathcal{S} = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$

Solve:

$$\min_{h \in \mathcal{H}} \left[ L_\mathcal{D}(h) = \mathbb{E}_{(x, y) \sim \mathcal{D}}[\ell(h, (x, y))] \right]$$

in the Probably (w.p. $1 - \delta$) Approximately Correct (up to $\epsilon$) sense
Statistical Learning Theory

Setting:
- The instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- The unknown probability distribution $\mathcal{D}$
- The hypotheses class $\mathcal{H}$
- The loss function $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$

Given: $\mathcal{S} = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$

Solve:
\[
\min_{h \in \mathcal{H}} \left[ L_{\mathcal{D}}(h) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \ell(h, (x, y)) \right] \right]
\]

in the Probably (w.p. $1 - \delta$) Approximately Correct (up to $\epsilon$) sense

Sample Complexity: $n(\delta, \epsilon) : (0, 1) \times (0, 1) \to \mathbb{N}$, the number of examples required to achieve $\epsilon$ accuracy with probability at least $1 - \delta$
Minimize the **EMPIRICAL** loss: 

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i))$$
Minimize the **EMPIRICAL** loss: 
\[ L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i)) \]

**Uniform Convergence:** If for any distribution \( \mathcal{D} \) over \( \mathcal{X} \) and for any sample \( S \) drawn i.i.d from \( \mathcal{D} \) it holds that for

\[ \forall h \in \mathcal{H}, \quad |L_S(h) - L_\mathcal{D}(h)| \leq \epsilon \]

**ERM with inductive bias**
- Restricting the \( \mathcal{H} \)
- Analytical properties of loss function \( \ell(\cdot, \cdot) \)
- Assumption on distribution \( \mathcal{D} \)
- Sparsity
- Margin
Empirical Risk Minimization (ERM)

- Minimize the **EMPIRICAL** loss: \( L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i)) \)

- **UNIFORM CONVERGENCE**: If for any distribution \( D \) over \( \mathcal{X} \) and for any sample \( S \) drawn i.i.d from \( D \) it holds that for

\[
\forall h \in \mathcal{H}, \quad |L_S(h) - L_D(h)| \leq \epsilon
\]

- **ERM with inductive bias**
  - ✓ Restricting the \( \mathcal{H} \)
  - ✓ Analytical properties of loss function \( \ell(\cdot, \cdot) \)
  - ✓ Assumption on distribution \( D \)
  - ✓ Sparsity
  - ✓ Margin

- **Fundamental Theorem of Learning Theory** [Vapnik and Chervonenkis, 1971]
Assumption:

The target risk $\epsilon$ is **known** to the learner!
Assumption: The target risk $\epsilon$ is known to the learner!

Question: Can we utilize this PRIOR KNOWLEDGE in the learning to improve the sample complexity?

Previous prior knowledges usually enter into the generalization bounds and have not been exploited in the learning process!
Known Lower/Upper Bounds

The Curse of Stochastic Oracle

Stochastic Gradient Descent with Target Risk
  Three Pillars
  SGD with Target Risk

Analysis

Conclusion and Future Work
Lower Bounds

\[ \Omega \left( \frac{1}{\epsilon} \left( \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right) \]

\[ \Omega \left( \frac{1}{\epsilon^2} \left( \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right) \]

[Ehrenfeucht et al., 1989; Blumer et al., 1989; Anthony and Bartlett, 1999]
Analytical properties of loss function (Smoothness and Strong Convexity) yield improved bounds:
Analytical properties of loss function (Smoothness and Strong Convexity) yield improved bounds:

**FAST RATES [Strong Convexity]**

\[ O \left( \frac{1}{\varepsilon} \left( \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right) \right) \]

[W. Lee and P. Bartlett (COLT’98), S. Kakade, A. Tewari (NIPS’08), S. Shalev-Shwartz, N. Srebro, K. Sridharan (NIPS’08)]

**OPTIMISTIC RATES [Smoothness]**

\[ O \left( \frac{1}{\varepsilon} \left( \frac{\varepsilon_{opt} + \varepsilon}{\varepsilon} \right) \left( \log^3 \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right) \right) \]

[N. Srebro, K. Sridharan, A. Tewari (NIPS’11)]
We assume that the learner is given the target expected risk in advance which we refer to as $\epsilon_{\text{prior}}$.

Surprisingly, we obtain an exponential improvement in the sample complexity:

$$O\left(dk^4 \left(\log \frac{1}{\epsilon_{\text{prior}}} \log \log \frac{1}{\epsilon_{\text{prior}}} + \log \frac{1}{\delta}\right)\right)$$

How?
Assumptions

Strong convexity:

\[ \ell(w_1) \geq \ell(w_2) + \langle \nabla \ell(w_2), w_1 - w_2 \rangle + \frac{\alpha}{2} \| w_1 - w_2 \|^2, \quad \forall \ w_1, w_2 \in \mathcal{H}. \]

Smoothness:

\[ \ell(w_1) \leq \ell(w_2) + \langle \nabla \ell(w_2), w_1 - w_2 \rangle + \frac{\beta}{2} \| w_1 - w_2 \|^2, \quad \forall \ w_1, w_2 \in \mathcal{H}. \]

Target risk assumption:

\[ \epsilon_{\text{prior}} \geq \epsilon_{\text{opt}} \]

Example: Regression with squared loss when the data matrix is not rank-deficient and \( \beta = \lambda_{\text{max}}(X^\top X) \)
Convex Learnability and The Curse of Stochastic Oracle
Not true in **Convex Learning Problems**!

[N. Srebro, O. Shamir, K. Sridharan (COLT’09, JMLR’11)]

Not true in **Multiclass Learning Problems**!

[A. Daniely, S. Sabato, S. Ben-David (COLT’11)]

Stochastic Convex Optimization $\iff$ Learnability in General Setting
ERM as Sample Average Approximation (SAA)

Alternatively, **directly** minimize the expected loss:

\[
\min_{w \in \mathcal{H}} \left[ L_{\mathcal{D}}(w) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \ell(w, (x,y)) \right] \right]
\]
ERMA SAA

Alternatively, directly minimize the expected loss:

\[
\min_{\mathbf{w} \in \mathcal{H}} \left[ L_D (\mathbf{w}) = \mathbb{E}_{(\mathbf{x}, y) \sim D} [\ell(\mathbf{w}, (\mathbf{x}, y))] \right]
\]

Stochastic Gradient Descent (SGD):

\[
\mathbf{w}_{t+1} = \Pi_{\mathcal{H}} (\mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t),
\]
Stochastic Optimization for Risk Minimization

ERM as Sample Average Approximation (SAA)

Alternatively, **directly** minimize the expected loss:

$$\min_{w \in \mathcal{H}} \left[ L_D(w) = \mathbb{E}_{(x,y) \sim D} [ \ell(w, (x,y))] \right]$$

Stochastic Gradient Descent (SGD):

$$w_{t+1} = \Pi_{\mathcal{H}} \left( w_t - \eta_t \hat{g}_t \right),$$

Stability as a necessary and sufficient condition for learnability

[S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan JMLR’11]

Lipschitzness or smoothness is necessary and boundedness and convexity alone are not sufficient!

∃ Stable AERM $\leftrightarrow$ Learnable with AERM $\leftrightarrow$ Learnable
Lower Bound for Stochastic Optimization

For any $\alpha$-strongly convex and $\beta$ smooth loss function and for any stochastic oracle with $E[\hat{g}] = \nabla L(w)$ and $E[\|\hat{g} - \nabla L(w)\|^2] \leq \sigma^2$, the following lower bond on the oracle complexity holds:

$$\mathcal{O}(1) \left( \sqrt{\frac{\beta}{\alpha}} \log \left( \frac{\beta \|w_0 - w_*\|^2}{\epsilon} \right) + \frac{\sigma^2}{\alpha \epsilon} \right).$$

[Nemirovski and Yudin, 1983]
Lower Bound for Stochastic Optimization

For any $\alpha$-strongly convex and $\beta$ smooth loss function and for any stochastic oracle with $\mathbb{E}[\hat{g}] = \nabla L(w)$ and $\mathbb{E}[\|\hat{g} - \nabla L(w)\|^2] \leq \sigma^2$, the following lower bond on the oracle complexity holds:

$$O(1) \left( \sqrt{\frac{\beta}{\alpha}} \log \left( \frac{\beta\|w_0 - w_*\|^2}{\epsilon} \right) + \frac{\sigma^2}{\alpha \epsilon} \right).$$

[Nemirovski and Yudin, 1983]

Life is easy if $\mathbb{E}[\|\hat{g} - \nabla L(w)\|^2] \approx O(\epsilon)!$ 🧐
Lower Bound for Stochastic Optimization

For any $\alpha$-strongly convex and $\beta$ smooth loss function and for any stochastic oracle with $\mathbb{E}[\hat{g}] = \nabla L(w)$ and $\mathbb{E}[\|\hat{g} - \nabla L(w)\|^2] \leq \sigma^2$, the following lower bond on the oracle complexity holds:

$$O(1) \left( \sqrt{\frac{\beta}{\alpha} \log \left( \frac{\|w_0 - w_*\|^2}{\epsilon} \right)} + \frac{\sigma^2}{\alpha \epsilon} \right).$$

[Nemirovski and Yudin, 1983]

▶ Life is easy if $\mathbb{E}\left[\|\hat{g} - \nabla L(w)\|^2\right] \approx O(\epsilon)!

▶ There is no control on the Stochastic Gradient Oracle!
Intuition: The Curse of Stochastic Oracle

Lower Bound for Stochastic Optimization

For any $\alpha$-strongly convex and $\beta$ smooth loss function and for any stochastic oracle with $E[\hat{g}] = \nabla L(w)$ and $E[\|\hat{g} - \nabla L(w)\|^2] \leq \sigma^2$, the following lower bond on the oracle complexity holds:

$$O(1) \left( \sqrt{\frac{\beta}{\alpha}} \log \left( \frac{\beta\|w_0 - w_*\|^2}{\epsilon} \right) + \frac{\sigma^2}{\alpha\epsilon} \right).$$

[Nemirovski and Yudin, 1983]

- **Life is easy if** $E[\|\hat{g} - \nabla L(w)\|^2] \approx O(\epsilon)$!

- **There is no control on the Stochastic Gradient Oracle!**

- **Solution**: Modify SGD to tolerate the noise in the gradients.
SGD with Target Risk
Three Pillars

Three main changes we have made to SGD:

- Run in **Multi-stages** with a **FIXED** size
- **Clip** the stochastic gradients
- **Shrink** the domain at each stage
Clipping the Stochastic Gradients

\[
[v_k^t]_i = \text{clip}(\gamma_k, [g_k^t]_i) = \text{sign}([g_k^t]_i) \min(\gamma_k, |[g_k^t]_i|)
\]
Clipping the Stochastic Gradients

\[
[v_k^t]_i = \text{clip}(\gamma_k, [g_k^t]_i) = \text{sign}([g_k^t]_i) \min(\gamma_k, |[g_k^t]_i|)
\]

Good news: reduces the variance
Bad news: unbiasedness of gradients no longer holds!

\[
E[v_k^t] \neq \nabla L(w_k^t) = E[g_k^t]
\]
At each stage $k$ we use a different hypothesis space $\mathcal{H}_k$ defined as:

$$\mathcal{H}_k = \{ \mathbf{w} \in \mathcal{H} : \| \mathbf{w} - \hat{\mathbf{w}}_k \| \leq \Delta_k \}$$

where $\Delta_{k+1} = \sqrt{\varepsilon \Delta_k^2 + \tau \epsilon_{\text{prior}}}$
Initialization: $\hat{w}_1 = 0$, $\Delta_1 = R$, and $\mathcal{H}_1 = \mathcal{H}$

for $k = 1, \ldots, m$

Set $w_{k} = \hat{w}_k$ and $\gamma_k = 2\xi\beta\Delta_k$
SGD with Target Risk

Initialization: \( \hat{w}_1 = 0, \Delta_1 = R, \) and \( H_1 = H \)

for \( k = 1, \ldots, m \)

Set \( w^t_k = \hat{w}_k \) and \( \gamma_k = 2\xi\beta\Delta_k \)

\[
\text{for } t = 1, \ldots, T_1 \\
\text{Receive training example } (x_t, y_t) \\
\text{Compute the gradient } \hat{g}^t_k 	ext{ and its clipped version } v^t_k \\
\text{Update the solution } w^{t+1}_k = \Pi_{H_k} (w^t_k - \eta v^t_k).
\]

end
SGD with Target Risk

**Initialization:** \( \hat{w}_1 = 0, \Delta_1 = R, \text{ and } \mathcal{H}_1 = \mathcal{H} \)

for \( k = 1, \ldots, m \)

Set \( w^t_k = \hat{w}_k \) and \( \gamma_k = 2\xi \beta \Delta_k \)

for \( t = 1, \ldots, T_1 \)

Receive training example \((x_t, y_t)\)

Compute the gradient \( g^t_k \) and its clipped version \( v^t_k \)

Update the solution \( w^{t+1}_k = \Pi_{\mathcal{H}_k} (w^t_k - \eta v^t_k) \).

end

Update \( \Delta_k \) as \( \Delta_{k+1} = \sqrt{\varepsilon \Delta_k^2 + \tau \epsilon_{\text{prior}}} \).

Compute the average solution \( \hat{w}_k = \sum_{t=1}^{T_1} \hat{w}^t_k / T_1 \)

Update the domain as \( \mathcal{H}_{k+1} = \{ w \in \mathcal{H} : \| w - \hat{w}_k \| \leq \Delta_{k+1} \} \)

end

Return \( \hat{w}_{m+1} \)
Convergence Rate

Assume that the hypothesis space $\mathcal{H}$ is compact and the loss function $\ell$ is $\alpha$-strongly convex and $\beta$-smooth, and $\epsilon_{\text{prior}}$ be the target expected loss given to the learner in advance such that $\epsilon_{\text{opt}} \leq \epsilon_{\text{prior}}$ holds. Then,

$$L(\hat{\mathbf{w}}_{m+1}) \leq \frac{\beta R^2}{2} \varepsilon^m + \left(1 + \frac{\tau}{1 - \varepsilon}\right) \epsilon_{\text{prior}},$$
If

\[ T \geq O \left( d \kappa^4 \left( \log \frac{1}{\epsilon_{\text{prior}}} \log \log \frac{1}{\epsilon_{\text{prior}}} + \log \frac{1}{\delta} \right) \right) \]

holds, then with a probability \( 1 - \delta \), the final solution \( \hat{w} \) attains a risk of \( O(\epsilon_{\text{prior}}) \), i.e., \( L(\hat{w}) \leq (1 + c)\epsilon_{\text{prior}} \).

\( \kappa = \beta/\alpha \) denotes the condition number of the loss function and \( d \) is the dimension of data.
Proof Sketch I

**Theorem 1**

For a fixed stage $k$, if $\|\hat{w}_k - w_*\| \leq \Delta_k$, then, with a probability $1 - \delta$, we have

$$\|\hat{w}_{k+1} - w_*\|^2 \leq a\Delta_k^2 + b \epsilon_{\text{prior}}$$

By the $\beta$-smoothness of $L(w)$, it implies that

$$L(\hat{w}_{m+1}) - L(w_*) \leq \frac{\beta}{2}\|\hat{w}_{m+1} - w_*\|^2 \leq \frac{\beta}{2} \epsilon_{\text{prior}} m \Delta_1^2 + \frac{\tau}{1 - \epsilon} \epsilon_{\text{prior}},$$

$$\leq \frac{\beta R^2}{2} \epsilon_{\text{prior}} m + \frac{\tau}{1 - \epsilon} \epsilon_{\text{prior}},$$
Key tools in proving the bound:

**Lemma 1: Deviation of a Clipped RV**

Let $X$ be a random variable, let $\tilde{X} = \text{clip}(X, C)$ and assume that $|\mathbb{E}[X]| \leq C/2$ for some $C > 0$. Then

$$|\mathbb{E}[\tilde{X}] - \mathbb{E}[X]| \leq \frac{2}{C} |\text{Var}[X]|$$

[E. Hazan and T. Koren (ICML'12)]

**Lemma 2: Self-boundedness of Smooth Functions**

For any $\beta$-smooth non-negative function $f : \mathbb{R} \to \mathbb{R}$, we have

$$|f'(w)| \leq \sqrt{4\beta f(w)}$$

[S. Shalev-Shwartz, Phd Thesis'07]

💡 Bernstein’s inequality for martingales

💡 Peeling process
Conclusions and Open Problems

Summary:

- We have studied passive learning with target risk as prior knowledge!
- We proposed modified SGD with three pillars: multi-staging, clipping, and shrinking which exploits the target risk in the learning
- We showed that the sample complexity is $\log \frac{1}{\epsilon_{\text{prior}}}$

Open Problems:

- Extension to non-parametric setting where hypotheses lie in a functional space of infinite dimension.
- Relation of target risk assumption we made to the low noise margin condition which is often made in active learning.

[Hanneke, 2009; Balcan et al., 2010]
Conclusions and Open Problems

Summary:

- We have studied passive learning with target risk as prior knowledge!
- We proposed modified SGD with three pillars: multi-staging, clipping, and shrinking which exploits the target risk in the learning.
- We showed that the sample complexity is $\log \frac{1}{\epsilon_{\text{prior}}}$.

Open Problems:

- Extension to non-parametric setting where hypotheses lie in a functional space of infinite dimension.
- Relation of target risk assumption we made to the low noise margin condition which is often made in active learning.
  
  [(Hanneke, 2009; Balcan et al., 2010]

- Improving the dependency on $d$ and the condition number $\kappa$. 
Thank you!