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 distibution $\mathcal{D}$
- The hypotheses class $\mathcal{H}$
- The loss function $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$

Statistical Learning Theory

Setting:

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

Given: $\mathcal{S} = ((\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)) \sim \mathcal{D}^n$

Solve:

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

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

Statistical Learning Theory

Setting:

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

Given: $\mathcal{S} = ((\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)) \sim \mathcal{D}^n$

Solve:

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

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

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

Empirical Risk Minimization (ERM)

☞ Minimize the EMPIRICAL loss: $L_{\mathcal{S}}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(h, (\mathbf{x}_i, y_i)\right)$

Empirical Risk Minimization (ERM)

☞ Minimize the EMPIRICAL loss: $L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h, (\mathbf{x}_i, y_i)\right)$

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

$$\forall h \in \mathcal{H}, \quad |L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

☞ 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: $L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h, (\mathbf{x}_i, y_i)\right)$

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

$$\forall h \in \mathcal{H}, \quad \left|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)\right| \leq \epsilon$$

☞ ERM with *inductive bias*

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

☞ Fundamental Theorem of Learning Theory [Vapnik and Chervonenkis, 1971]

$$\begin{tikzcd}
\boxed{\text{Finite Dim.}} \arrow[r, Rightarrow] & \boxed{\text{Uniform Convergence}} \arrow[r, Rightarrow] & \boxed{\text{Learnable with ERM}} \arrow[r, Rightarrow] & \boxed{\text{Learnable}} \arrow[lll, Rightarrow, bend left]
\end{tikzcd}$$

Property Testing of Learning

Assumption:

The target risk ϵ is **known** to the learner!

Property Testing of Learning

Assumption:

The target risk ϵ 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!

Outline

Lower Bounds

☞ PAC Setting

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

☞ AGNOSTIC PAC Setting

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

[Ehrenfeucht et al., 1989; Blumer et al., 1989; Anthony and Bartlett, 1999]

Fast and Optimistic Sample Complexities

☞ Analytical properties of loss function (Smoothness and Strong Convexity) yield improved bounds:

Fast and Optimistic Sample Complexities

☞ Analytical properties of loss function (Smoothness and Strong Convexity) yield improved bounds:

☞ FAST RATES [Strorng Convexity]

$$O\left(\frac{1}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

[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]

$$O\left(\frac{1}{\epsilon}\left(\frac{\epsilon_{\text{opt}} + \epsilon}{\epsilon}\right)\left(\log^3\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

[N. Srebro, K. Sridharan, A. Tewari (NIPS'11)]

Main Result on Sample Complexity

☞ We assume that the learner is given the target expected risk in advance which we refer to as ϵ_{prior}

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

$$\mathcal{O}\left(d\kappa^4\left(\log\frac{1}{\epsilon_{\mathrm{prior}}}\log\log\frac{1}{\epsilon_{\mathrm{prior}}}+\log\frac{1}{\delta}\right)\right)$$

☞ How?

Assumptions

☞ **Strong convexity:**

$$\ell(\mathbf{w}_1) \geq \ell(\mathbf{w}_2) + \langle \nabla \ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\alpha}{2} \|\mathbf{w}_1 - \mathbf{w}_2\|^2, \ \forall \ \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{H}.$$

☞ **Smoothness:**

$$\ell(\mathbf{w}_1) \leq \ell(\mathbf{w}_2) + \langle \nabla \ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\beta}{2} \|\mathbf{w}_1 - \mathbf{w}_2\|^2, \ \forall \ \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{H}.$$

☞ **Target risk assumption**:

$$\epsilon_{\text{prior}} \geq \epsilon_{\text{opt}}$$

Example: Regression with squared loss when the data matrix is not rank-deficient and $\beta = \lambda_{\max}(X^\top X)$

Convex Learnability
and
The Curse of Stochastic Oracle

Learning without Uniform Convergence

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 $\Longleftrightarrow$ Learnability in General Setting

Stochastic Optimization for Risk Minimization

- ☞ ERM as Sample Average Approximation (SAA)
- ☞ Alternatively, directly minimize the expected loss:

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

Stochastic Optimization for Risk Minimization

- ☞ ERM as Sample Average Approximation (SAA)
- ☞ Alternatively, directly minimize the expected loss:

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

- ☞ Stochastic Gradient Descent (SGD):

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

Stochastic Optimization for Risk Minimization

☞ ERM as Sample Average Approximation (SAA)

☞ Alternatively, directly minimize the expected loss:

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

☞ Stochastic Gradient Descent (SGD):

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

☞ 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!

$\exists$ Stable AERM $\Longleftrightarrow$ Learnable with AERM $\Longleftrightarrow$ Learnable

Intuition: The Curse of Stochastic Oracle

Lower Bound for Stochastic Optimization

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

$$\mathcal{O}(1)\left(\sqrt{\frac{\beta}{\alpha}}\log\left(\frac{\beta\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\epsilon}\right) + \frac{\sigma^2}{\alpha\epsilon}\right).$$

[Nemirovski and Yudin, 1983]

Intuition: The Curse of Stochastic Oracle

Lower Bound for Stochastic Optimization

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

$$\mathcal{O}(1)\left(\sqrt{\frac{\beta}{\alpha}}\log\left(\frac{\beta\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\epsilon}\right) + \frac{\sigma^2}{\alpha\epsilon}\right).$$

[Nemirovski and Yudin, 1983]

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

Intuition: The Curse of Stochastic Oracle

Lower Bound for Stochastic Optimization

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

$$\mathcal{O}(1)\left(\sqrt{\frac{\beta}{\alpha}}\log\left(\frac{\beta\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\epsilon}\right) + \frac{\sigma^2}{\alpha\epsilon}\right).$$

[Nemirovski and Yudin, 1983]

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

➤ There is no control on the Stochastic Gradient Oracle!

Intuition: The Curse of Stochastic Oracle

Lower Bound for Stochastic Optimization

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

$$\mathcal{O}(1)\left(\sqrt{\frac{\beta}{\alpha}}\log\left(\frac{\beta\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\epsilon}\right) + \frac{\sigma^2}{\alpha\epsilon}\right).$$

[Nemirovski and Yudin, 1983]

- Life is easy if $\mathbb{E}\left[\|\hat{\mathbf{g}} - \nabla L(\mathbf{w})\|^2\right] \approx O(\epsilon)$!
- There is no control on the Stochastic Gradient Oracle!
- Solution: Modify SGD to tolerate the noise in the gradients.

SGD with Target Risk

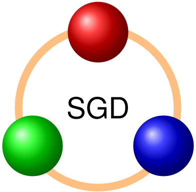
Three Pillars

Three main changes we have made to SGD:

☞ Run in Multi-stages with a FIXED size

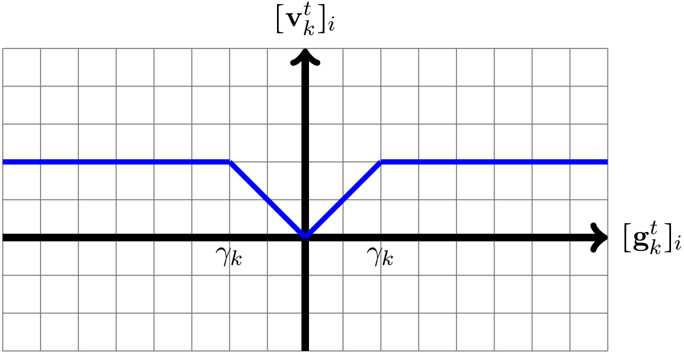
☞ Clip the stochastic gradients

☞ Shrink the domain at each stage



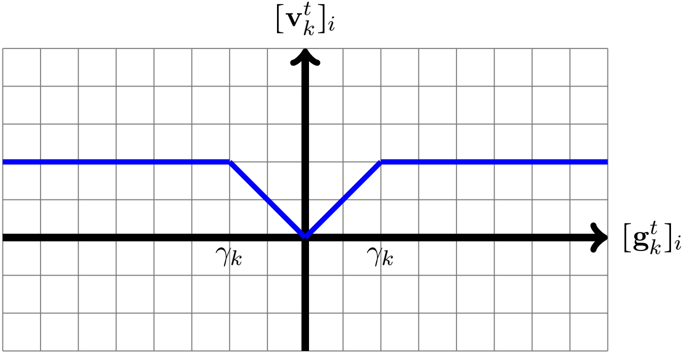
Clipping the Stochastic Gradients

$$[\mathbf{v}_k^t]_i = \mathrm{clip}(\gamma_k, [\mathbf{g}_k^t]_i) = \mathrm{sign}([\mathbf{g}_k^t]_i)\min(\gamma_k, \left|[\mathbf{g}_k^t]_i\right|)$$



Clipping the Stochastic Gradients

$$[\mathbf{v}_k^t]_i = \mathrm{clip}(\gamma_k, [\mathbf{g}_k^t]_i) = \mathrm{sign}([\mathbf{g}_k^t]_i) \min(\gamma_k, \left|[\mathbf{g}_k^t]_i\right|)$$



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

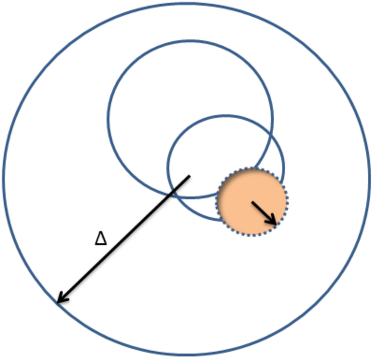
$$\mathbf{E}[\mathbf{v}_k^t] \neq \left[\nabla L(\mathbf{w}_k^t) = \mathbf{E}[\mathbf{g}_k^t]\right]$$

Shrinking the Hypothesis Space $\mathcal{H}$

At each stage k we use a different hypothesis space $\mathcal{H}_k$ defined as:

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

where $\Delta_{k+1} = \sqrt{\varepsilon \Delta_k^2 + \tau \epsilon_{\text{prior}}}$



SGD with Target Risk

Initialization: $\widehat{\mathbf{w}}_1 = 0$, $\Delta_1 = R$, and $\mathcal{H}_1 = \mathcal{H}$

for $k = 1, \ldots, m$ [Epoch]

Set $\mathbf{w}_k^t = \widehat{\mathbf{w}}_k$ and $\gamma_k = 2\xi\beta\Delta_k$

SGD with Target Risk

Initialization: $\widehat{\mathbf{w}}_1 = 0$, $\Delta_1 = R$, and $\mathcal{H}_1 = \mathcal{H}$

for $k = 1, \ldots, m$ [Epoch]

Set $\mathbf{w}_k^t = \widehat{\mathbf{w}}_k$ and $\gamma_k = 2\xi\beta\Delta_k$

for $t = 1, \ldots, T_1$ [SGD]

Receive training example $(\mathbf{x}_t, y_t)$

Compute the gradient $\hat{\mathbf{g}}_k^t$ and its clipped version $\mathbf{v}_k^t$

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

end

SGD with Target Risk

Initialization: $\widehat{\mathbf{w}}_1 = 0$, $\Delta_1 = R$, and $\mathcal{H}_1 = \mathcal{H}$

for $k = 1, \ldots, m$ [Epoch]

Set $\mathbf{w}_k^t = \widehat{\mathbf{w}}_k$ and $\gamma_k = 2\xi\beta\Delta_k$

for $t = 1, \ldots, T_1$ [SGD]

Receive training example $(\mathbf{x}_t, y_t)$

Compute the gradient $\hat{\mathbf{g}}_k^t$ and its clipped version $\mathbf{v}_k^t$

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

end

Update Δ_k as $\Delta_{k+1} = \sqrt{\varepsilon \Delta_k^2 + \tau \epsilon_{\text{prior}}}$. [Shrinking]

Compute the average solution $\widehat{\mathbf{w}}_k = \sum_{t=1}^{T_1} \widehat{\mathbf{w}}_k^t / T_1$

Update the domain as $\mathcal{H}_{k+1} = \{\mathbf{w} \in \mathcal{H} : \|\mathbf{w} - \widehat{\mathbf{w}}_k\| \leq \Delta_{k+1}\}$

end

Return $\widehat{\mathbf{w}}_{m+1}$

Convergence Rate

Convergence Rate

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

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

Sample Complexity

If

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

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

☞ $\kappa = \beta/\alpha$ 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 $\|\widehat{\mathbf{w}}_k - \mathbf{w}_*\| \leq \Delta_k$, then, with a probability $1 - \delta$, we have

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

By the β-smoothness of $L(\mathbf{w})$, it implies that

$$\begin{aligned} L(\widehat{\mathbf{w}}_{m+1}) - L(\mathbf{w}_*) \leq \frac{\beta}{2}\|\widehat{\mathbf{w}}_{m+1} - \mathbf{w}_*\|^2 &\leq \frac{\beta}{2}\varepsilon^m \Delta_1^2 + \frac{\tau}{1-\varepsilon}\epsilon_{\text{prior}}, \\ &\leq \frac{\beta R^2}{2}\varepsilon^m + \frac{\tau}{1-\varepsilon}\epsilon_{\text{prior}}, \end{aligned}$$

Proof Sketch II

Key tools in proving the bound:

Lemma 1: Deviation of a Clipped RV

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

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

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

Lemma 2: Self-boundedness of Smooth Functions

For any β-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 $\log \frac{1}{\epsilon_{\mathsf{prior}}}$

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 κ


Thank you