Basic Concepts of Online Learning

Mathematical Analysis of Machine Learning Algorithms
(Chapter 14)

Online Learning Model

The online learning learning model can be considered as a repeated game. For $t = 1, 2, \ldots,$

- An adversary picks (X_t, Y_t), and reveals X_t only.
- An online learning algorithm $\mathcal{A}$ predicts $\hat{f}_{t-1}(X_t)$.
- The value of Y_t is revealed and a loss $L(\hat{f}_{t-1}(X_t), Y_t)$ is computed.

The goal of online learning is to minimize the aggregated loss

$$\sum_{t=1}^{T} L(\hat{f}_{t-1}(X_t), Y_t).$$

Regret Bound

An online algorithm $\mathcal{A}$ considers model class $\mathcal{F} = \{f(w, x) : w \in \Omega\}$, and learn $w_{t-1} \in \Omega$ at time t based on previously observed data $\mathcal{S}_{t-1} = \{Z_1, \ldots, Z_{t-1}\}$. It makes prediction on next observation:

$$\hat{f}_{t-1}(X_t) = f(w_{t-1}, X_t) \qquad \text{with } w_{t-1} = \mathcal{A}(\mathcal{S}_{t-1}).$$

The performance of online learning is measured by regret bound, which is analogous to oracle inequality in supervised learning.

Regret

The aggregated loss of an online algorithm is compared to the optimal aggregated loss in hindsight:

$$\text{REG}_T = \sum_{t=1}^{T} L(f(w_{t-1}, X_t), Y_t) - \inf_{w \in \Omega} \sum_{t=1}^{T} L(f(w, X_t), Y_t) \leq \epsilon_T. \qquad (1)$$

Online Binary Classification with Perceptron

Consider the binary classification problem with $Y \in \{\pm 1\}$, and linear functions

$$f(w, X) = w^\top X,$$

with prediction rule:

$$\begin{cases} 1 & f(w, X) \geq 0 \\ -1 & \text{otherwise} \end{cases}.$$

The loss function is binary classification error: $\mathbb{1}(f(w, X)Y \leq 0)$.

The Perceptron algorithm is an algorithm which works with data coming in sequentially as in Algorithm 1. It is *mistake-driven*, which means it only updates the model weight vector when the prediction makes a mistake.

Perceptron Algorithm

Algorithm 1: Perceptron Algorithm

Input: Sequence $(X_1, Y_1), \ldots, (X_T, Y_T)$

Output: w_s

Let $w_0 = 0$

for $t = 1, 2, \ldots, T$ **do**

    Observe X_t and predict label $\operatorname{sign}(w_{t-1}^\top X_t)$

    Observe Y_t and compute mistake $\mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0)$

    if $w_{t-1}^\top X_t Y_t > 0$ **then**

        // No mistake

        Let $w_t = w_{t-1}$

    else

        // A mistake is observed

        Let $w_t = w_{t-1} + X_t Y_t$

Randomly pick s from 0 to $T - 1$

Return: w_s

Perceptron Mistake (Regret) Bound

Theorem 1 (Thm 14.1)

Consider the perceptron Algorithm in Algorithm 1. Consider $\gamma > 0$ *and weight vector* w_* *such that for all* t

$$w_*^\top X_t Y_t \geq \gamma.$$

Then we have the following mistake bound:

$$\sum_{t=1}^{T} \mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0) \leq \frac{\|w_*\|_2^2 \sup\{\|X_t\|_2^2\}}{\gamma^2}.$$

Proof of Theorem 1

Let $M = \sup_t \|X_t\|_2$, and let $\eta = \gamma/M^2$. Assume that we have a mistake at time step t, then we have

$$(\eta w_{t-1} - w_*)^\top X_t Y_t \leq 0 - w_*^\top X_t Y_t \leq -\gamma.$$

This implies that

$$\begin{aligned}
\|\eta w_t - w_*\|_2^2 =& \|\eta w_{t-1} + \eta X_t Y_t - w_*\|_2^2 \\
=& \|\eta w_{t-1} - w_*\|_2^2 + 2\eta(\eta w_{t-1} - w_*)^\top X_t Y_t + \eta^2 \|X_t\|_2^2 \\
\leq& \|\eta w_{t-1} - w_*\|_2^2 - 2\eta\gamma + \eta^2 M^2 \\
\leq& \|\eta w_{t-1} - w_*\|_2^2 - \frac{\gamma^2}{M^2}.
\end{aligned}$$

Note also that $\|\eta w_t - w_*\|_2^2 = \|\eta w_{t-1} - w_*\|_2^2$ if there is no mistake at time step t. Therefore by summing over $t = 1$ to $t = t$, we obtain

$$0 \leq \|\eta w_t - w_*\|_2^2 \leq \|\eta w_0 - w_*\|_2^2 - \frac{\gamma^2}{M^2} k,$$

where k is the number of mistakes. This implies the bound.

Multi-class Perceptron

For multi-class prediction with q classes $y \in \{1, \ldots, q\}$, we may use the notations of vector prediction functions in the analysis of kernel methods.
Consider a vector prediction function $f(x) \in \mathbb{R}^q$, with linear prediction model for class ℓ defined as:

$$f_\ell(x) = w^\top \psi(x, \ell).$$

The predicted class for each x is

$$\hat{y}(w, x) \in \arg\max_\ell \; w^\top \psi(x, \ell),$$

and the error (or mistake) for an instance x with true label y is

$$\mathbb{1}(\hat{y}(w, x) \neq y).$$

Multiclass Perceptron Algorithm

Algorithm 2: Multi-Class Perceptron Algorithm

Input: Sequence $(X_1, Y_1), \ldots, (X_T, Y_T)$

Output: w_s

Let $w_0 = 0$

for $t = 1, 2, \ldots, T$ **do**

  Observe X_t and predict label $\hat{Y}_t \in \arg\max_\ell \{w_{t-1}^\top \psi(X_t, \ell)\}$

  Observe Y_t and compute mistake $\mathbb{1}(\hat{Y}_t \neq Y_t)$

  if $\hat{Y}_t == Y_t$ **then**

    // No mistake

    Let $w_t = w_{t-1}$

  else

    // A mistake is observed

    Let $w_t = w_{t-1} + [\psi(X_t, Y_t) - \psi(X_t, \hat{Y}_t)]$

Randomly pick s from 0 to $T - 1$

Return: w_s

Mistake Bound

Theorem 2 (Thm 14.2)

Consider Algorithm 2. We have the following mistake bound:

$$\sum_{t=1}^{T} \mathbb{1}(\hat{Y}_t \neq Y_t) \leq \inf_{\gamma>0, w} \left[\sum_{t=1}^{T} 2 \max\left(0, 1 - \gamma^{-1} \min_{\ell \neq Y_t} w^\top [\psi(X_t, Y_t) - \psi(X_t, \ell)]\right) + \frac{\|w\|_2^2 \sup\{\|\psi(X_t, Y_t) - \psi(X_t, Y_\ell)\|_2^2\}}{\gamma^2} \right].$$

Proof of Theorem 2 (I/II)

The proof is basically the same as that of the binary case. Given any $\gamma > 0$ and w. We let $\psi_t = \psi(X_t, Y_t) - \psi(X_t, \hat{Y}_t)$, $M = \sup\{\|\psi_t\|_2\}$, and $\eta = \gamma / M^2$. Assume that we have a mistake at time step t, then we have $\hat{Y}_t \neq Y_t$, and $w_{t-1}^\top \psi_t \leq 0$. It implies that

$$(\eta w_{t-1} - w_*)^\top \psi_t \leq 0 - w_*^\top \psi_t \leq \max(0, \gamma - w_*^\top \psi_t) - \gamma.$$

Therefore by taking

$$\begin{aligned}
\|\eta w_t - w_*\|_2^2 =& \|\eta w_{t-1} + \eta \psi_t - w_*\|_2^2 \\
=& \|\eta w_{t-1} - w_*\|_2^2 + 2\eta(\eta w_{t-1} - w_*)^\top \psi_t + \eta^2 \|\psi_t\|_2^2 \\
\leq& \|\eta w_{t-1} - w_*\|_2^2 + 2\eta \max(0, \gamma - w_*^\top \psi_t) - 2\eta\gamma + \eta^2 M^2 \\
\leq& \|\eta w_{t-1} - w_*\|_2^2 + 2\eta \max(0, \gamma - w_*^\top \psi_t) - \frac{\gamma^2}{M^2}.
\end{aligned}$$

Proof of Theorem 2 (II/II)

Note also that $\|\eta w_t - w_*\|_2^2 = \|\eta w_{t-1} - w_*\|_2^2$ if there is no mistake at time step t. Therefore by summing over $t = 1$ to $t = T$, we obtain

$$0 \leq \|\eta w_T - w_*\|_2^2 \leq 2\eta \sum_{t=1}^{T} \max(0, \gamma - w_*^\top \psi_t) + \|\eta w_0 - w_*\|_2^2 - \frac{\gamma^2}{M^2} k,$$

where k is the number of mistakes. This implies the bound.

Online to Batch Conversion

Assume that in online learning, the observed data are random, with $Z_t = (X_t, Y_t) \sim \mathcal{D}$. Assume also that we have a regret bound of the general form:

$$\sum_{t=1}^{T} \phi(w_{t-1}, Z_t) \leq \epsilon(S_T). \tag{2}$$

By taking expectations, we obtain an expected generalization bound

$$\mathbb{E}_{S_T} \sum_{t=1}^{T} \mathbb{E}_Z \phi(w_{t-1}, Z) \leq \mathbb{E}_{S_T} \epsilon(S_T).$$

If we select s uniformly from 0 to $T-1$, then

$$\mathbb{E}_{S_T} \mathbb{E}_s \mathbb{E}_Z \phi(w_s, Z) \leq \mathbb{E}_{S_T} \epsilon(S_T)/T.$$

This leads to an oracle inequality in the batch (supervised learning) setting.

Expected Oracle Inequality Example

For the perceptron algorithm, we may let

$$\phi(w, Z) = \mathbb{1}(\hat{y}(w, X) \neq Y) - 2 \max\left(0, 1 - \gamma^{-1} \min_{\ell \neq Y} w^\top [\psi(X, Y) - \psi(X, \ell)]\right).$$

Proposition 3

Consider Algorithm 2, with s chosen uniformly at random from 0 to $T-1$. If $Z_t = (X_t, Y_t) \sim \mathcal{D}$ are iid observations, then we have

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}_{T-1}} \mathbb{E}_s \mathbb{E}_{Z \sim \mathcal{D}} \mathbb{1}(\hat{y}(w_s, X) \neq Y) \\ &\leq \inf_{\gamma > 0, w} \left[2\mathbb{E}_{Z \sim \mathcal{D}} \max\left(0, 1 - \gamma^{-1} \min_{\ell \neq Y_t} w^\top [\psi(X, Y) - \psi(X, \ell)]\right) \right. \\ &\qquad \left. + \frac{\|w\|_2^2 \sup\{\|\psi(X, Y) - \psi(X, Y')\|_2^2\}}{\gamma^2 \, T} \right], \end{aligned}$$

where the prediction rule is $\hat{y}(w, x) \in \arg\max_\ell w^\top \psi(x, \ell)$.

High Probability Bound

We can use martingale tail probability bounds in the online learning analysis. Assume that we have high probability result of the following form: with probability at least $1 - \delta$ over $\mathcal{S}_T$,

$$\sum_{t=1}^{T} \mathbb{E}_{Z \sim \mathcal{D}} \phi(w_{t-1}, Z) \leq \sum_{t=1}^{T} \phi(w_{t-1}, Z_t) + \epsilon(\delta). \tag{3}$$

We may combine this bound with (2), and obtain the following probability bound for the randomized estimator s, uniformly chosen from 0 to $T - 1$. With probability at least $1 - \delta$ over $\mathcal{S}_T$:

$$\mathbb{E}_s \, \mathbb{E}_{Z \sim \mathcal{D}} \phi(w_s, Z) \leq \frac{\epsilon(\delta) + \epsilon(\mathcal{S}_T)}{T}.$$

Example: Perceptron Algorithm

Proposition 4

Consider Algorithm 1, with s uniformly drawn from 0 to $T-1$. Assume $w_*^\top XY \geq \gamma > 0$ for all $Z = (X, Y)$. If $Z_t = (X_t, Y_t) \sim \mathcal{D}$, then with probability at least $1-\delta$:

$$\mathbb{E}_s \mathbb{E}_{Z \in \mathcal{D}} \mathbb{1}(\hat{y}(w_s, X) \neq Y)$$
$$\leq \inf_{\lambda > 0} \left[\frac{\lambda}{1 - e^{-\lambda}} \frac{\|w_*\|_2^2 \sup_X \|X\|_2^2}{\gamma^2 T} + \frac{\ln(1/\delta)}{(1 - e^{-\lambda})\ T} \right].$$

Proof of Proposition 4

Let

$$\{\xi_i = \mathbb{1}(w_{i-1}^\top X_i Y_i \leq 0) : i = 1, 2, \ldots, n\}$$

be a sequence of random variables, Theorem 13.5 implies that for any $\lambda > 0$, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{(X_i,Y_i)\sim\mathcal{D}}\xi_i \leq \frac{\lambda}{1-e^{-\lambda}}\frac{1}{T}\sum_{i=1}^{T}\xi_i + \frac{\ln(1/\delta)}{(1-e^{-\lambda})\ T}.$$

Also note that the mistake bound in Theorem 1 implies that

$$\sum_{i=1}^{T}\xi_i \leq \frac{\|w_*\|_2^2 \sup\{\|X\|_2^2\}}{\gamma^2}.$$

Since $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\mathbb{1}(w_{i-1}^\top XY \leq 0) = \mathbb{E}_{(X_i,Y_i)\sim\mathcal{D}}\xi_i$, we obtain the desired result.

Online Convex Optimization

One can extend the analysis of the Perceptron algorithms to general convex loss functions, leading to the so-called online convex optimization.

Algorithm 3: Online Gradient Descent

Input: Sequence of loss functions $\ell_1, \ldots, \ell_T$ defined on Ω
Output: $\hat{w}$

Let $w_0 = 0$
for $t = 1, 2, \ldots, T$ **do**
    Observe loss $\ell_t(w_{t-1})$
    Let $\tilde{w}_t = w_{t-1} - \eta_t \nabla \ell_t(w_{t-1})$
    Let $w_t = \arg\min_{w \in \Omega} \|w - \tilde{w}_t\|_2^2$
Let $\hat{w} = T^{-1} \sum_{t=1}^{T} w_{t-1}$ or $\hat{w} = w_s$ for a random s from 0 to $T-1$
Return: $\hat{w}$

Regret Bound

Theorem 5 (Thm 14.5)

Let $\{\ell_t(w) : w \in \Omega\}$ be a sequence of real-valued convex loss functions defined on a convex set Ω. Assume that all $\ell_t(w)$ are G-Lipschitz (that is, $\|\nabla \ell_t(w)\|_2 \leq G$). If we let $\eta_t = \eta > 0$ be a constant in Algorithm 3. Then for all $w \in \Omega$, we have

$$\sum_{t=1}^{T} \ell_t(w_{t-1}) \leq \sum_{t=1}^{T} \ell_t(w) + \frac{\|w_0 - w\|_2^2}{2\eta} + \frac{\eta T}{2} G^2.$$

Proof of Theorem 5 (I/II)

We have the following inequality:

$$\begin{aligned}
\|\tilde{w}_t - w\|_2^2 =& \|w_{t-1} - w - \eta \nabla \ell_t(w_{t-1})\|_2^2 \\
=& \|w_{t-1} - w\|_2^2 - 2\eta \nabla \ell_t(w_{t-1})^\top (w_{t-1} - w) + \eta^2 \|\nabla \ell_t(w_{t-1})\|_2^2 \\
\leq& \|w_{t-1} - w\|_2^2 - 2\eta \nabla \ell_t(w_{t-1})^\top (w_{t-1} - w) + G^2 \eta^2 \\
\leq& \|w_{t-1} - w\|_2^2 - 2\eta [\ell_t(w_{t-1}) - \ell_t(w)] + G^2 \eta^2,
\end{aligned}$$

where the first inequality used the Lipschitz condition, and the second inequality used the convexity condition.

Proof of Theorem 5 (II/II)

Since $w_t \in \Omega$ is the projection of $\tilde{w}_t$ onto Ω and $w \in \Omega$. We also have

$$\|w_t - w\|_2^2 \leq \|\tilde{w}_t - w\|_2^2.$$

Therefore, we have

$$\|w_t - w\|_2^2 \leq \|w_{t-1} - w\|_2^2 - 2\eta[\ell_t(w_{t-1}) - \ell_t(w)] + G^2\eta^2.$$

Now we may sum over $t = 1$ to $t = T$, and obtain

$$\|w_T - w\|_2^2 \leq \|w_0 - w\|_2^2 - 2\eta \sum_{t=1}^{T} [\ell_t(w_{t-1}) - \ell_t(w)] + TG^2\eta^2.$$

Rearrange the terms, we obtain the desired bound.

A More General Result

The following result applies both for online convex optimization and perceptron algorithm.

Theorem 6 (Thm 14.6)

Consider Algorithm 5 with the update rule replaced by the following method

$$\tilde{w}_t = w_{t-1} - \eta_t g_t.$$

If we can choose g_t so that

$$g_t^\top (w - w_{t-1}) \leq \tilde{\ell}_t(w) - \ell_t(w_{t-1}),$$

then

$$\sum_{t=1}^{T} \ell_t(w_{t-1}) \leq \sum_{t=1}^{T} \tilde{\ell}_t(w) + \frac{\|w_0 - w\|_2^2}{2\eta} + \frac{\eta T}{2} G^2.$$

Example

Example 7

When $w_{t-1}^\top X_t Y_t \leq 0$, we have

$$\begin{aligned}-(w - w_{t-1})^\top X_t Y_t \leq & \gamma - w^\top X_t Y_t - \gamma \\ & \leq \max(0, \gamma - w^\top X_t Y_t) - \gamma \mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0).\end{aligned}$$

When $w_{t-1}^\top X_t Y_t > 0$, we have

$$0 \leq \max(0, \gamma - w^\top X_t Y_t) - \gamma \mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0).$$

Therefore let $g_t = -\mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0) X_t Y_t$, then

$$(w - w_{t-1})^\top g_t \leq \max(0, \gamma - w^\top X_t Y_t) - \gamma \mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0).$$

This implies that Theorem 1 is a special case of Theorem 6 by taking $\tilde{\ell}_t(w) = \max(0, \gamma - w^\top X_t Y_t)$ and $\ell_t(w_{t-1}) = \gamma \mathbb{1}(w_{t-1}^\top X_t Y_t \leq 0)$.

Oracle Inequality

Theorem 8 (Thm 14.9)

Consider loss function $\phi(w, Z) \in [0, M]$ with $Z \sim \mathcal{D}$, and $w \in \Omega$, where Ω is a convex set. Assume that $\phi(w, Z)$ is convex and G-Lipschitz with respect to w. Let $[Z_1, \ldots, Z_T] \sim \mathcal{D}^T$ be independent samples, and consider $\hat{w}$ obtained from Algorithm 3, with $\ell_i(w) = \phi(w, Z_i)$ and $\eta_t = \eta > 0$. Then with probability at least $1 - \delta$,

$$\mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{w}, Z) \leq \inf_{w\in\Omega}\left[\mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) + \frac{1}{2\eta T}\|w - w_0\|_2^2\right] + \frac{\eta}{2}G + M\sqrt{\frac{2\ln(2/\delta)}{T}}.$$

Proof of Theorem 8 (I/II)

Note that the convexity and Jensen's inequality implies that

$$\mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{w}, Z) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_t}\phi(w_{t-1}, Z_t). \tag{4}$$

Moreover, using the Azuma's inequality, we have with probability at least $1-\delta/2$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_t}\phi(w_{t-1}, Z_t) \leq \frac{1}{T}\sum_{t=1}^{T}\phi(w_{t-1}, Z_t) + M\sqrt{\frac{\ln(2/\delta)}{2T}}. \tag{5}$$

Using Theorem 5, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\phi(w_{t-1}, Z_t) \leq \frac{1}{T}\sum_{t=1}^{T}\phi(w, Z_t) + \frac{\|w_0 - w\|_2^2}{2\eta T} + \frac{\eta}{2}G^2. \tag{6}$$

Using the Chernoff bound, we have with probability at least $1-\delta/2$:

$$\frac{1}{T}\sum_{t=1}^{T}\phi(w, Z_t) \leq \mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) + M\sqrt{\frac{\ln(2/\delta)}{2T}}. \tag{7}$$

Proof of Theorem 8 (II/II)

By taking the union bound, and combine the above four inequalities, we obtain the following. With probability at least $1-\delta$:

$$\begin{aligned}
\mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{w},Z) \leq & \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_t}\phi(w_{t-1},Z_t) \\
\leq & \frac{1}{T}\sum_{t=1}^{T}\phi(w_{t-1},Z_t) + M\sqrt{\frac{\ln(2/\delta)}{2T}} \\
\leq & \frac{1}{T}\sum_{t=1}^{T}\phi(w,Z_t) + \frac{\|w_0-w\|_2^2}{2\eta T} + \frac{\eta}{2}G^2 + M\sqrt{\frac{\ln(2/\delta)}{2T}} \\
\leq & \mathbb{E}_{Z\sim\mathcal{D}}\phi(w,Z) + \frac{\|w_0-w\|_2^2}{2\eta T} + \frac{\eta}{2}G^2 + M\sqrt{\frac{2\ln(2/\delta)}{T}}.
\end{aligned}$$

The first inequality used (4). The second inequality used (5). The third inequality used (6). The last inequality used (7).

Compare with ERM

If we take $\eta = O(1/\sqrt{T})$, then we obtain a convergence result of $O(1/\sqrt{T})$ in Theorem 8.

In the ERM oracle inequality for kernel methods using Rademacher complexity analysis, the loss function does not have to be convex. We have the following result based on Corollary 9.27:

$$\mathbb{E}_{\mathcal{D}}\phi(\hat{w}, Z) \leq \mathbb{E}_{\mathcal{D}}\phi(w, Z) + \lambda\|w\|_2^2 + O\left(\sqrt{\frac{\ln((\lambda + B^2)/(\delta\lambda))}{n}}\right) + O\left(\frac{B^2 \ln((\lambda + B^2)/(\delta\lambda))}{\lambda n}\right).$$

This result is similar to that of Theorem 8 with $\eta = 1/(\lambda T)$.

Example

Example 9

Consider the structured-SVM loss of Example 9.32, where

$$\phi(w, z) = \max_{\ell}[\gamma(y, \ell) - w^\top(\psi(x, y) - \psi(x, \ell))].$$

If $\|\psi(x, y) - \psi(, \ell)\|_2 \leq B$, then we take $G = B$ in Theorem 5:

$$\frac{1}{T}\sum_{t=1}^{T}\phi(w_{t-1}, Z_t) \leq \frac{1}{T}\sum_{t=1}^{T}\phi(w, Z_t) + \frac{\|w_0 - w\|_2^2}{2\eta T} + \frac{\eta}{2}B^2.$$

Taking expectation, we obtain with $\lambda = 1/(\eta T)$

$$\mathbb{E}_{S_T}\,\mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{w}, Z) \leq \mathbb{E}_{Z\sim\mathcal{D}}\,\phi(w, Z) + \frac{\lambda}{2}\|w_0 - w\|_2^2 + \frac{1}{2\lambda T}B^2.$$

Independent of class size q, better than Rademacher complexity result in Example 9.32.

Strong Convexity

For L_2 regularization (or kernel methods), one can obtain a better bound using strong convexity.

Definition 10

A convex function $\ell(w)$ is λ strongly convex for some $\lambda > 0$ if

$$\ell(w') \geq \ell(w) + \nabla \ell(w)^\top (w' - w) + \frac{\lambda}{2} \|w' - w\|_2^2.$$

Observe that for regularized loss, we take

$$\ell_t(w) = \phi(w, Z_t) + \frac{\lambda}{2} \|w - w_0\|_2^2. \tag{8}$$

If $\phi(w, z)$ is convex in w, then $\ell_t(w)$ is λ strongly convex.

Regret Bound

The following result holds for strongly convex loss functions, and the learning rate has been proposed to solve SVMs

Theorem 11 (Thm 14.11)

Consider convex loss functions $\ell_t(w) : \Omega \to \mathbb{R}$*, which are G-Lipschitz (that is,* $\|\nabla \ell_t(w)\|_2 \leq G$*) and* λ *strongly convex. If we let* $\eta_t = 1/(\lambda t) > 0$ *in Algorithm 3, then for for all* $w \in \Omega$*, we have*

$$\sum_{t=1}^{T} \ell_t(w_{t-1}) \leq \sum_{t=1}^{T} \ell_t(w) + \frac{1 + \ln T}{2\lambda} G^2.$$

Proof of Theorem 11 (I/II)

Similar to the proof of Theorem 5, we have

$$
\begin{aligned}
&\|w_t - w\|_2^2 \\
\leq &\|\tilde{w}_t - w\|_2^2 \\
= &\|w_{t-1} - w - \eta_t \nabla \ell_t(w_{t-1})\|_2^2 \\
= &\|w_{t-1} - w\|_2^2 - 2\eta_t \nabla \ell_t(w_{t-1})^\top (w_{t-1} - w) + \eta_t^2 \|\nabla \ell_t(w_{t-1})\|_2^2 \\
\leq &\|w_{t-1} - w\|_2^2 + 2\eta_t [\ell_t(w) - \ell_t(w_{t-1})] - \eta_t \lambda \|w_{t-1} - w\|_2^2 + G^2 \eta_t^2 \\
= &(1 - \lambda \eta_t) \|w_{t-1} - w\|_2^2 + 2\eta_t [\ell_t(w) - \ell_t(w_{t-1})] + G^2 \eta_t^2,
\end{aligned}
$$

where strong-convexity is used to derive the second inequality.

Proof of Theorem 11 (II/II)

Note that $1 - \eta_t\lambda = \eta_t/\eta_{t-1}$ and for notation convenience we take $1/\eta_0 = 0$. This implies that

$$\eta_t^{-1}\|w_t - w\|_2^2 \leq \eta_{t-1}^{-1}\|w_{t-1} - w\|_2^2 + 2[\ell_t(w) - \ell_t(w_{t-1})] + G^2\eta_t.$$

By summing over $t = 1$ to $t = T$, we obtain

$$\eta_T^{-1}\|w_T - w\|_2^2 \leq \eta_0^{-1}\|w_0 - w\|_2^2 + 2\sum_{t=1}^{T}[\ell_t(w) - \ell_t(w_{t-1})] + G^2\sum_{t=1}^{T}\frac{1}{\lambda t}.$$

Using $\sum_{t=1}^{T}(1/t) \leq 1 + \ln T$, we obtain the desired bound.

Weighted Regret

It is possible to remove the $\ln T$ factor if we use weighted regret.

Theorem 12 (Thm 14.12)

Consider convex loss functions $\ell_t(w) : \Omega \to \mathbb{R}$, which are G-Lipschitz (that is, $\|\nabla \ell_t(w)\|_2 \leq G$) and λ strongly convex. If we let $\eta_t = 2/(\lambda(t+1)) > 0$ in Algorithm 3, then for for all $w \in \Omega$, we have

$$\sum_{t=1}^{T} \frac{2(t+1)}{T(T+3)} \ell_t(w_{t-1}) \leq \sum_{t=1}^{T} \frac{2(t+1)}{T(T+3)} \ell_t(w) + \frac{2G^2}{\lambda(T+3)}.$$

Proof of Theorem 12

As in the proof of Theorem 11, we have

$$\|w_t - w\|_2^2 \leq (1 - \eta_t \lambda)\|w_{t-1} - w\|_2^2 + 2\eta_t[\ell_t(w) - \ell_t(w_{t-1})] + G^2\eta_t^2.$$

This implies that $\eta_t^{-2}(1 - \eta_t\lambda) \leq \eta_{t-1}^{-2}$, where we set $\eta_0^{-2} = 0$:

$$\eta_t^{-2}\|w_t - w\|_2^2 \leq \eta_{t-1}^{-2}\|w_{t-1} - w\|_2^2 + 2\eta_t^{-1}[\ell_t(w) - \ell_t(w_{t-1})] + G^2.$$

By summing over $t = 1$ to $t = T$, we obtain

$$\eta_T^{-2}\|w_T - w\|_2^2 \leq \eta_0^{-2}\|w_0 - w\|_2^2 + 2\sum_{t=1}^{T} \eta_t^{-1}[\ell_t(w) - \ell_t(w_{t-1})] + G^2 T.$$

This leads to the bound.

Oracle Inequality

Corollary 13

Consider the regularized loss function (8) *with* $w_0 = 0$*, where* $\phi(w, z)$ *is convex in* w*, and* G *Lipschitz in* w*. Moreover assume that* $d(\Omega) = \sup\{\|w\|_2 : w \in \Omega\}$*. If* $Z_1, \ldots, Z_T \sim \mathcal{D}$ *are independent samples, then we can obtain the following expected oracle inequality for Algorithm 3 if we take learning rate in Theorem 11:*

$$\mathbb{E}_{S_T}\mathbb{E}_{\mathcal{D}}\phi(\hat{w}, Z) + \frac{\lambda}{2}\|\hat{w}\|_2^2 \leq \inf_{w\in\Omega}\left[\mathbb{E}_{\mathcal{D}}\phi(w, Z) + \frac{\lambda}{2}\|w\|_2^2\right] + \frac{\ln(eT)}{2\lambda T}[G + \lambda d(\Omega)]^2.$$

We can also obtain the following expected oracle inequality for Algorithm 3 if we take learning rate in Theorem 12 with $\hat{w}' = \sum_{t=1}^{T} \frac{2(t+1)}{T^2+3T} w_{t-1}$*, then*

$$\mathbb{E}_{S_T}\mathbb{E}_{\mathcal{D}}\phi(\hat{w}', Z) + \frac{\lambda}{2}\|\hat{w}'\|_2^2 \leq \inf_{w\in\Omega}\left[\mathbb{E}_{\mathcal{D}}\phi(w, Z) + \frac{\lambda}{2}\|w\|_2^2\right] + \frac{2[G + \lambda d(\Omega)]^2}{\lambda(T+3)}.$$

Hedge Algorithm for Nonconvex Problem

Algorithm 4: Hedge Algorithm

Input: T, prior $p_0(w)$ on Ω, learning rate $\eta > 0$

Randomly draw $w_0 \sim p_0(w)$

for $t = 1, 2, \ldots, T$ **do**

  Observe loss $\ell_t(w_{t-1})$

  Randomly draw $w_t \sim p_t(w)$ according to

$$p_t(w) \propto p_0(w) \exp\left(-\eta \sum_{s=1}^{t} \ell_s(w)\right), \tag{9}$$

  where $p_0(w)$ is a prior on Ω.

Regret Bound

Theorem 14 (Thm 14.15)

Assume that for all t:

$$\sup_{w \in \Omega} \ell_t(w) - \inf_{w \in \Omega} \ell_t(w) \leq M,$$

then Algorithm 4 has regret

$$\sum_{t=1}^{T} \mathbb{E}_{w_{t-1} \sim p_{t-1}(\cdot)} \ell_t(w_{t-1}) \leq \inf_{p \in \Delta(\Omega)} \left[\mathbb{E}_{w \sim p} \sum_{t=1}^{T} \ell_t(w) + \frac{1}{\eta} \mathrm{KL}(p || p_0) \right] + \frac{\eta T M^2}{8},$$

where $\Delta(\Omega)$ denotes the set of probability distributions on Ω.

Result used in the Proof of Theorem 14

Proposition 15 (Prop 7.16)

Given any function $U(w)$, we have

$$\min_{p \in \Delta(\Omega)} [\mathbb{E}_{w \sim p} U(w) + \mathrm{KL}(p||p_0)] = -\ln \mathbb{E}_{w \sim p_0} \exp(-U(w)),$$

and the solution is achieved by the Gibbs distribution

$$q(w) \propto p_0(w) \exp(-U(w)).$$

Here $\Delta(\Omega)$ denotes the set of probability distributions on Ω.

Proof of Theorem 14

Let

$$Z_t = -\ln \mathbb{E}_{w\sim p_0} \exp\left(-\eta \sum_{s=1}^{t} \ell_s(w)\right)$$

be the log-partition function for observations up to time t. We have

$$\begin{aligned} Z_{t-1} - Z_t &= \ln \mathbb{E}_{w\sim p_{t-1}} \exp\left(-\eta \ell_t(w)\right) \\ &\leq -\eta \mathbb{E}_{w\sim p_{t-1}} \ell_t(w) + \frac{\eta^2 M^2}{8}, \end{aligned}$$

where the first equation is simple algebra, and the inequality follows from the estimate of logarithmic moment generation function in Lemma 2.15. By summing over $t = 1$ to T, and noticing that $Z_0 = 0$, we obtain

$$\sum_{t=1}^{T} \mathbb{E}_{w_{t-1}\sim p_{t-1}(\cdot)} \ell_t(w_{t-1}) \leq \frac{1}{\eta} Z_T + \frac{\eta T M^2}{8}.$$

The desired bound follows by applying Proposition 15 to reformulate the log-partition function Z_T.

Example

If Ω contains a discrete number of functions, and consider p to be a measure concentrated on a single $w \in \Omega$, then $\mathrm{KL}(p||p_0) = \ln(1/p_0(w))$. We thus obtain from Theorem 14

$$\sum_{t=1}^{T} \mathbb{E}_{w_{t-1} \sim p_{t-1}(\cdot)} \ell_t(w_{t-1}) \leq \inf_{w \in \Omega} \left[\sum_{t=1}^{T} \ell_t(w) + \frac{1}{\eta} \ln \frac{1}{p_0(w)} \right] + \frac{\eta T M^2}{8}.$$

If $|\Omega| = N$ with $p_0(w) = 1/N$, then by setting $\eta = \sqrt{8 \ln N/(TM^2)}$, we obtain

$$\sum_{t=1}^{T} \mathbb{E}_{w_{t-1} \sim p_{t-1}(\cdot)} \ell_t(w_{t-1}) \leq \inf_{w \in \Omega} \sum_{t=1}^{T} \ell_t(w) + M \sqrt{\frac{T \ln N}{2}}.$$

This matches the generalization result using empirical process in Chapter 3. Large probability bounds can be obtained by using online to batch conversion with Azuma's inequality.

Summary (Chapter 14)

- Basics of Online Learning
- Perceptron Mistake Bounds
- Online to Batch Conversion
- Online Convex Optimization
  - first order gradient algorithm
  - non-strongly-convex regret bound
  - strongly-convex regret bound
- Online nonconvex optimization