Online Aggregation and Second Order Algorithms

Mathematical Analysis of Machine Learning Algorithms
(Chapter 15)

Log-Loss for Density Estimation

Consider conditional density estimation with log-loss (negative-log-likelihood), where the loss function is

$$\phi(w, Z) = -\ln p(Y|w, X).$$

Example 1

For discrete $y \in \{1, \ldots, K\}$, we have

$$p(y|w, x) = \frac{\exp(f_y(w, x))}{\sum_{k=1}^{K} \exp(f_k(w, x))},$$

and (let $Z = (x, y)$):

$$\phi(w, Z) = -f_y(w, x) + \ln \sum_{k=1}^{K} \exp(f_k(w, x)).$$

Example 2 (More Log-Loss Example)

For least squares regression with noise variance σ^2, we may have

$$p(y|w,x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-f(w,x))^2}{2\sigma^2}\right),$$

and (let $Z=(x,y)$):

$$\phi(w,Z) = \frac{(y-f(w,x))^2}{2\sigma^2} + \ln(\sqrt{2\pi}\sigma).$$

We may also consider the noise as part of the model parameter, and let

$$p(y|[w,\sigma],x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-f(w,x))^2}{2\sigma^2}\right)$$
$$\phi([w,\sigma],Z) = \frac{(y-f(w,x))^2}{2\sigma^2} + \ln(\sqrt{2\pi}\sigma).$$

Bayesian Posterior Averaging

Bayesian Posterior Distribution

Consider a prior $p_0(w)$ on Ω. Given the training data $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$, the posterior distribution is

$$p(w|\mathcal{S}_n) = \frac{\prod_{i=1}^n p(Y_i|w, X_i)p_0(w)}{\int_\Omega \prod_{i=1}^n p(Y_i|w', X_i)p_0(w')\, dw'}. \quad (1)$$

The *Bayesian posterior average estimator* is the averaged probability estimate over the posterior

$$\hat{p}(y|x, \mathcal{S}_n) = \int_\Omega p(y|w, x)p(w|\mathcal{S}_n)\, dw. \quad (2)$$

However, we do not assume that the Bayesian assumption holds true in the theoretical analysis.

Regret Bound: Property of Log-Partition Function

Proposition 3 (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.

Regret Bound: Conditional Density Estimation

Theorem 4 (Thm 15.3)

We have

$$
\begin{aligned}
-\sum_{t=1}^{T} \ln \hat{p}(Y_t, |X_t, \mathcal{S}_{t-1}) &= -\ln \mathbb{E}_{w\sim p_0} \prod_{t=1}^{T} p(Y_t|w, X_t) \\
= \quad &\inf_{q\in\Delta(\Omega)} \left[-\mathbb{E}_{w\sim q} \sum_{t=1}^{T} \ln p(Y_t|w, X_t) + \mathbb{E}_{w\sim q} \ln \frac{q(w)}{p_0(w)} \right],
\end{aligned}
$$

where $\Delta(\Omega)$ is the set of probability distributions over Ω.

Proof of Theorem 4

We have

$$
\begin{aligned}
\ln \hat{p}(Y_t, |X_t, \mathcal{S}_{t-1}) &= \ln \int_\Omega p(Y_t|w, X_t) p(w|\mathcal{S}_{t-1}) \, dw \\
&= \ln \frac{\int_\Omega \prod_{i=1}^{t} p(Y_i|w, X_i) p_0(w) dw}{\int_\Omega \prod_{i=1}^{t-1} p(Y_i|w', X_i) p_0(w') dw'} \\
&= \ln \mathbb{E}_{w \in p_0} \prod_{i=1}^{t} p(Y_i|w, X_i) - \ln \mathbb{E}_{w \in p_0} \prod_{i=1}^{t-1} p(Y_i|w, X_i).
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{t=1}^{T} \ln \hat{p}(Y_t, |x_t, \mathcal{S}_{t-1}) &= \ln \mathbb{E}_{w \in p_0} \prod_{t=1}^{T} p(Y_t|w, X_t) \\
&= \sup_q \left[\mathbb{E}_{w \sim q} \ln \prod_{t=1}^{T} p(Y_t|w, X_t) - \mathbb{E}_{w \sim q} \ln \frac{q(w)}{p_0(w)} \right],
\end{aligned}
$$

where the second equality used Proposition 3.

Discrete Family of Probability Distributions

Corollary 5 (Cor 15.4)

If $\Omega = \{w_1, \ldots\}$ *is discrete, then*

$$-\sum_{t=1}^{T} \ln \hat{p}(Y_t, |X_t, \mathcal{S}_{t-1}) \leq \inf_{w \in \Omega} \left[-\sum_{t=1}^{T} \ln p(Y_t | w, X_t) - \ln p_0(w) \right].$$

Proof.

Given any $w' \in \Omega$, if we choose $q(w) = 1$ when $w = w'$, and $q(w) = 0$ when $w \neq w'$, then from Theorem 4:

$$-\mathbb{E}_{w \sim q} \sum_{t=1}^{T} \ln p(Y_t | w, X_t) + \mathbb{E}_{w \sim q} \ln \frac{q(w)}{p_0(w)} = -\sum_{t=1}^{T} \ln p(Y_t | w', X_t) - \ln p_0(w').$$

□

Example: Finite $|\Omega| = N$

Let $p_0(w) = 1/N$ be the uniform distribution on Ω, then we have

$$-\sum_{t=1}^{T} \ln \hat{p}(Y_t, |X_t, \mathcal{S}_{t-1}) \leq \inf_{w \in \Omega} \left[-\sum_{t=1}^{T} \ln p(Y_t | w, X_t) + \ln N \right].$$

This means that we have a constant regret which is independent of T.

Using online to batch conversion:

$$-\frac{1}{T} \mathbb{E}_{\mathcal{S}_T} \sum_{t=1}^{T} \mathbb{E}_{Z \sim \mathcal{D}} \ln \hat{p}(Y | X, \mathcal{S}_{t-1}) \leq \inf_{w \in \Omega} \left[-\mathbb{E}_{Z \sim \mathcal{D}} \ln p(Y | w, X) + \frac{\ln N}{T} \right].$$

Ridge Regression

The general regret bound for Bayesian model averaging can be used to analyze the ridge regression method. Consider the following linear prediction problem with least squares loss:

$$f(w, x) = w^\top \psi(x),$$

with loss function

$$(y - f(w, x))^2.$$

Consider the following ridge regression estimator:

$$\hat{w}(\mathcal{S}_n) = \arg\min_w \left[\sum_{i=1}^{n} (Y_i - w^\top \psi(X_i))^2 + \|w\|_{\Lambda_0}^2, \right], \tag{3}$$

where Λ_0 is a symmetric positive definite matrix, which is often chosen as λI for some $\lambda > 0$ in applications.

Bayesian Interpretation of Ridge Regression

Proposition 6 (Prop 15.5)

Consider probability model $p(y|w, x) = N(w^\top \psi(x), \sigma^2)$*, with prior* $p_0(w) = N(0, \sigma^2 \Lambda_0^{-1})$*. Then given* $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$*, we have*

$$p(w|\mathcal{S}_n) = N(\hat{w}(\mathcal{S}_n), \sigma^2 \hat{\Lambda}(\mathcal{S}_n)^{-1}),$$

where $\hat{w}(\mathcal{S}_n)$ *is given by* (3) *and*

$$\hat{\Lambda}(\mathcal{S}_n) = \sum_{i=1}^{n} \psi(X_i)\psi(X_i)^\top + \Lambda_0.$$

Given x*, the posterior distribution of* y *is*

$$\hat{p}(y|x, \mathcal{S}_n) = N\left(\hat{w}(\mathcal{S}_n)^\top \psi(x), \sigma^2 + \sigma^2 \psi(x)^\top \hat{\Lambda}(\mathcal{S}_n)^{-1} \psi(x)\right).$$

The result holds even when the Bayesian model isn't correct.

Proof of Proposition 6 (I/II)

It is clear that

$$p(w|\mathcal{S}_n) \propto \exp\left(-\sum_{i=1}^{n} \frac{(w^\top \psi(X_i) - Y_i)^2}{2\sigma^2} - \frac{\|w\|_{\Lambda_0}^2}{2\sigma^2}\right).$$

Note that (3) implies that

$$\begin{aligned} &\sum_{i=1}^{n} \frac{(w^\top \psi(X_i) - Y_i)^2}{2\sigma^2} + \frac{\|w\|_{\Lambda_0}^2}{2\sigma^2} \\ =&\sum_{i=1}^{n} \frac{(\hat{w}^\top \psi(X_i) - Y_i)^2}{2\sigma^2} + \frac{\|\hat{w}\|_{\Lambda_0}^2}{2\sigma^2} + \frac{1}{2\sigma^2}(w - \hat{w})^\top \hat{\Lambda}(\mathcal{S}_n)(w - \hat{w}). \end{aligned} \quad (4)$$

Proof of Proposition 6 (II/II)

This implies the first desired result. Moreover, given x, and let the random variable $u = w^\top \psi(x)$ with $w \sim p(w|\mathcal{S}_n)$, we have

$$u|x, \mathcal{S}_n \sim N(\hat{w}^\top \psi(x), \sigma^2 \psi(x)^\top \hat{\Lambda}(\mathcal{S}_n)^{-1} \psi(x)).$$

Since in posterior distribution, the observation $y|u \sim u + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$, we know that

$$y|x, \mathcal{S}_n \sim N(\hat{w}^\top \psi(x), \sigma^2 \psi(x)^\top \hat{\Lambda}^{-1} \psi(x) + \sigma^2).$$

This implies the desired result.

Predictive Loss Bound

Theorem 7 (Thm 15.6)

Consider the ridge regression method of (3)*. We have the following result for any* $\sigma \geq 0$ *and for all observed sequence* $\mathcal{S}_T$*:*

$$\sum_{t=1}^{T} \left[\frac{(Y_t - \hat{w}(\mathcal{S}_{t-1})^\top \psi(X_t))^2}{b_t} + \sigma^2 \ln b_t \right]$$

$$= \inf_{w} \left[\sum_{t=1}^{T} (Y_t - w^\top \psi(X_t))^2 + \|w\|_{\Lambda_0}^2 \right] + \sigma^2 \ln \left| \Lambda_0^{-1} \Lambda_T \right|,$$

where

$$\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} \psi(X_s)\psi(X_s)^\top,$$

and $b_t = 1 + \psi(X_t)^\top \Lambda_{t-1}^{-1} \psi(X_t)$*.*

Proof of Theorem 7 (I/II)

Assume $w \in \mathbb{R}^d$. We note that from Gaussian integration that

$$\mathbb{E}_{w\sim p_0} \prod_{i=1}^{T} p(Y_i|w, X_i)$$

$$= \int \frac{|\Lambda_0^{-1}|^{-1/2}}{(2\pi)^{(T+d)/2}\sigma^T\sigma^d} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(Y_i - w^\top\psi(X_i))^2 - \frac{\|w\|_{\Lambda_0}^2}{2\sigma^2}\right) dw$$

$$= \frac{|\Lambda_0^{-1}\Lambda_T|^{-1/2}}{(2\pi)^{T/2}\sigma^T} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(Y_i - \hat{w}(\mathcal{S}_T)^\top\psi(X_i))^2 - \frac{\|\hat{w}(\mathcal{S}_T)\|_{\Lambda_0}^2}{2\sigma^2}\right),$$

where the last equation used Gaussian integration with decomposition (4).

Proof of Theorem 7 (II/II)

That is,

$$-\ln \mathbb{E}_{w\sim p_0} \prod_{i=1}^{T} p(Y_i|w, X_i) = T\ln(\sqrt{2\pi}\sigma) + \frac{1}{2}\ln\left|\Lambda_0^{-1}\Lambda_T\right| + \frac{1}{2\sigma^2}\sum_{i=1}^{T}(Y_i - \hat{w}(\mathcal{S}_T)^\top \psi(X_i))^2 + \frac{\|\hat{w}(\mathcal{S}_T)\|_{\Lambda_0}^2}{2\sigma^2}.$$

Moreover, from Proposition 6, we have

$$\sum_{i=1}^{T} -\ln p(Y_t|\hat{w}(\mathcal{S}_{t-1}), X_t) = \sum_{t=1}^{T}\left[\frac{(Y_t - \hat{w}(\mathcal{S}_{t-1})^\top \psi(X_t))^2}{2\sigma^2 b_t} + \frac{1}{2}\ln(b_t) + \ln(\sqrt{2\pi}\sigma)\right].$$

The desired result now follows from Theorem 4.

Regret Bound with Bounded Response

Corollary 8 (Cor 15.7)

Assume that $Y_t \in [0, M]$ for $t \geq 1$. Consider the ridge regression method of (3)*, and let*

$$\hat{Y}_t = \max(0, \min(M, \hat{w}(\mathcal{S}_{t-1})^\top \psi(X_t))).$$

We have

$$\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 \leq \inf_{w} \left[\sum_{t=1}^{T} (Y_t - w^\top \psi(X_t))^2 + \|w\|_{\Lambda_0}^2 \right] + M^2 \ln \left| \Lambda_0^{-1} \Lambda_T \right|,$$

where

$$\Lambda_T = \Lambda_0 + \sum_{s=1}^{T} \psi(X_s) \psi(X_s)^\top.$$

Proof of Corollary 8

We can apply Theorem 7 by taking $\sigma^2 = M^2$. By using the following inequality

$$0 \leq \frac{1 - b_t}{b_t} + \ln b_t,$$

we obtain

$$\begin{aligned}
\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 &\leq \sum_{t=1}^{T} \left[(Y_t - \hat{Y}_t)^2 + M^2 \frac{1 - b_t}{b_t} + M^2 \ln b_t \right] \\
&\leq \sum_{t=1}^{T} \left[(Y_t - \hat{Y}_t)^2 + (Y_t - \hat{Y}_t)^2 \frac{1 - b_t}{b_t} + M^2 \ln b_t \right] \qquad (5) \\
&\leq \sum_{t=1}^{T} \left[\frac{(Y_t - \hat{w}(\mathcal{S}_{t-1})^\top \psi(X_i))^2}{b_t} + M^2 \ln b_t \right],
\end{aligned}$$

where b_t is defined in Theorem 7. Note that (5) used $1 - b_t \leq 0$ and $(Y_t - \hat{Y}_t)^2 \leq M^2$. The desired result is now a direct application of Theorem 7.

Estimation of Determinant

Proposition 9 (Simplification of Prop 15.8)

Given any $\mathcal{X}$ and $\psi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is an inner product space. Then for each $\lambda > 0$ and integer T, the embedding entropy of $\psi(\cdot)$ can be defined as

$$\mathrm{entro}(\lambda, \psi(\mathcal{X})) = \sup_{\mathcal{D}} \ln \left| I + \frac{1}{\lambda} \mathbb{E}_{X \sim \mathcal{D}} \psi(X) \psi(X)^\top \right| .$$

If $\sup_{X \in \mathcal{X}} \|\psi(X)\|_{\mathcal{H}} \leq B$ *and* $\dim(\mathcal{H}) < \infty$*, then*

$$\mathrm{entro}(\lambda, \psi(\mathcal{X})) \leq \dim(\mathcal{H}) \ln \left(1 + \frac{B^2}{\dim(\mathcal{H}) \lambda} \right) .$$

One can also deal with the situation of $\dim(\mathcal{H}) = \infty$ (see Proposition 15.8).

Proof of Proposition 9

Let $A = I + (\lambda)^{-1}\mathbb{E}_{X\sim\mathcal{D}}\psi(X)\psi(X)^\top$ and $d = \mathrm{dim}(\mathcal{H})$, then

$$\mathrm{trace}(A) \leq d + (\lambda)^{-1}\mathbb{E}_{X\sim\mathcal{D}}\mathrm{trace}(\psi(X)\psi(X)^\top) \leq d + B^2/\lambda.$$

Using the AM-GM inequality, we have

$$|A| \leq [\mathrm{trace}(A)/d]^d \leq (1 + B^2/(d\lambda))^d.$$

Example

Example 10

Consider Corollary 8 with $\Lambda_0 = \lambda I$. We can use Proposition 9 to obtain

$$\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2$$
$$\leq \inf_{w}\left[\sum_{t=1}^{T}(Y_t - w^\top \psi(X_t))^2 + \lambda \|w\|_2^2\right] + M^2 d \ln\left(1 + \frac{TB^2}{d\lambda}\right),$$

where we assume that $\|\psi(x)\|_2 \leq B$, and d is the dimension of $\psi(x)$.

Online to Batch Conversion

Assume $(X_t, Y_t) \sim \mathcal{D}$ are iid examples. By taking expectation with respect to $\mathcal{D}$, and by using Jensen's inequality for the concave log-determinant function, we obtain (see Corollary 15.11) that with $\Lambda_0 = \lambda I$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_T} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{X \sim \mathcal{D}} \left(f_*(X) - \hat{f}(\hat{w}(\mathcal{S}_{t-1}), X) \right)^2 \\
&\leq \inf_w \mathbb{E}_{X \sim \mathcal{D}} \left[(f_*(X) - w^\top \psi(X))^2 + \lambda \|w\|_2^2 \right] \\
&\qquad + \frac{M^2 + \sigma^2}{T} \ln \left| I + \frac{T}{\lambda} \mathbb{E}_{X \sim \mathcal{D}} \psi(X) \psi(X)^\top \right| .
\end{aligned}
$$

Note that this bound is superior to the Rademacher complexity bound, and the best convergence rate can be achieved is $O(\ln T/T)$.

Exponential Model Aggregation

Consider a general loss function with $Z = (X, Y)$:

$$\phi(w, Z) = L(f(w, X), Y),$$

where $L(f, y)$ is convex with respect to f. Consider a prior $p_0(w)$ on Ω, and the following form of Gibbs distribution (which we will also refer to as posterior):

$$p(w|\mathcal{S}_n) \propto \exp\left[-\eta \sum_{i=1}^{n} \phi(w, Z_i)\right] p_0(w), \tag{6}$$

where $\eta > 0$ is a learning rate parameter. The exponential model aggregation algorithm computes

$$\hat{f}(x|\mathcal{S}_n) = \int_\Omega f(w, x) p(w|\mathcal{S}_n)\, dw, \tag{7}$$

where $p(w|\mathcal{S}_n)$ is given by (6).

Online Exponential Model Aggregation

Algorithm 1: Online Exponential Model Aggregation

Input: $\eta > 0$, $\{f(w, x) : w \in \Omega\}$, prior $p_0(w)$

Output: $\hat{f}(\cdot|\mathcal{S}_T)$

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

  Observe X_t

  Let $\hat{f}_t = \hat{f}(X_t|\mathcal{S}_{t-1})$ according to (7)

  Observe Y_t

  Compute $L(\hat{f}_t, Y_t)$

Return: $\hat{f}(\cdot|\mathcal{S}_T)$

Exponential Concavity

In order to analyze Algorithm 1, we need to employ the concept of α-exponential concavity introduced below.

Definition 11 (Def 15.12)

A convex function $g(u)$ is α-exponential concave for some $\alpha > 0$ if

$$e^{-\alpha g(u)}$$

is concave in u.

Properties

Proposition 12 (Prop 15.13)

A convex function $\phi(u)$ is α exponentially concave if

$$\alpha\nabla\phi(u)\nabla\phi(u)^\top \leq \nabla^2\phi(u).$$

Proof.

We have

$$\nabla^2 e^{-\alpha\phi(u)} = e^{-\alpha\phi(u)}\left[-\alpha\nabla^2\phi(u) + \alpha^2\nabla\phi(u)\nabla\phi(u)^\top\right] \leq 0.$$

This implies the concavity of $\exp(-\alpha\phi(u))$. □

Examples

Example 13

We note that if $\phi(u)$ is both Lipschitz $\|\nabla\phi(u)\|_2 \leq G$, and λ-strongly convex, then

$$(\lambda/G^2)\nabla\phi(u)\nabla\phi(u)^\top \leq \lambda I \leq \nabla^2\phi(u).$$

Proposition 12 implies that $\phi(u)$ is λ/G^2 exponentially concave.

Examples (cont)

Example 14

Consider the loss function $L(u, y) = (u - y)^2$. If $|u - y| \leq M$, then $L(u, y)$ is α-exponentially concave in u with $\alpha \leq 1/(2M^2)$.

Example 15

Consider a function $f(\cdot)$, and let $L(f(\cdot), y) = -\ln f(y)$, then $L(f(\cdot), y)$ is α exponentially concave in $f(\cdot)$ for $\alpha \leq 1$. This loss function is applicable to conditional probability estimate $\ln f(y|x)$.

Regret Bound

Theorem 16 (Thm 15.19)

Assume that $L(f, y)$ is η-exponentially concave. Then (7) satisfies the following regret bound:

$$\sum_{t=1}^{T} L(\hat{f}(X_t|\mathcal{S}_{t-1}), Y_t) \leq \inf_q \left[\mathbb{E}_{w\sim q} \sum_{t=1}^{T} L(f(w, X_t), Y_t) + \frac{1}{\eta} \mathbb{E}_{w\sim q} \ln \frac{q(w)}{p_0(w)} \right].$$

We note that Theorem 4 is a special case of Theorem 16, with $\eta = 1$,

$$L(f(w, x), y) = -\ln P(y|w, x)$$

and $f(w, x) = P(y|w, x)$. In this case, $\exp(-L(f, y)) = f_y$ is concave in f.

Proof of Theorem 16

Since $e^{-\eta L(f,y)}$ is concave in f, we obtain from Jensen's inequality

$$\ln \int e^{-\eta L(f(w,x),y)} p(w|\mathcal{S}_{t-1}) dw \leq \ln e^{-\eta L(\hat{f}(x|\mathcal{S}_{t-1}),y)}.$$

With $(x, y) = (X_t, Y_t)$, this can be equivalently rewritten as

$$L(\hat{f}(X_t|\mathcal{S}_{t-1}), Y_t) \leq \frac{-1}{\eta} \ln \frac{\int_\Omega \exp(-\eta \sum_{i=1}^{t} L(f(w, X_i), Y_i)) p_0(w) dw}{\int_\Omega \exp(-\eta \sum_{i=1}^{t-1} L(f(w, X_i), Y_i)) p_0(w) dw}.$$

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

$$\sum_{t=1}^{T} L(\hat{f}(X_t|\mathcal{S}_{t-1}), Y_t) \leq \frac{-1}{\eta} \ln \int_\Omega \exp\left(-\eta \sum_{i=1}^{T} L(f(w, X_i), Y_i)\right) p_0(w) dw.$$

Using Proposition 3, we obtain the desired result.

Example: Log-Loss

Example 17

Theorem 4 is a special case of Theorem 16, with $\eta = 1$,

$$L(f(\cdot|w,x),y) = -\ln P(y|w,x)$$

and $f(\cdot|w,x) = P(\cdot|w,x)$. In this case,

$$\exp(-L(f(\cdot|\cdot),y)) = f(y|\cdot)$$

is a component of $f(\cdot|\cdot)$ indexed by y, and thus concave in $f(\cdot|\cdot)$.

Example: Least Squares

Example 18

Assume that $L(f, y) = (f - y)^2$, and $\sup |f(w, x) - y| \leq M$. Then for $\eta \leq 1/(2M^2)$, $L(f, y)$ is η exponentially concave. Therefore we have

$$\sum_{t=1}^{T} (\hat{f}(X_t|\mathcal{S}_{t-1}) - Y_t)^2 \leq \inf_q \left[\mathbb{E}_{w \sim q} \sum_{t=1}^{T} (f(w, X_t) - Y_t)^2 + \frac{1}{\eta} \mathbb{E}_{w \sim q} \ln \frac{q(w)}{p_0(w)} \right].$$

In particular, if Ω is countable, then

$$\sum_{t=1}^{T} (\hat{f}(X_t|\mathcal{S}_{t-1}) - Y_t)^2 \leq \inf_{w \in \Omega} \left[\sum_{t=1}^{T} (f(w, X_t) - Y_t)^2 + \frac{1}{\eta} \ln \frac{1}{p_0(w)} \right].$$

Model aggregation is superior to ERM for misspecified models, because the regret with respect to the best function in the function class is still $O(1/n)$.

Adaptive Gradient

Algorithm 2: Adaptive SubGradient Method (AdaGrad)

Input: $\eta > 0$, w_0, A_0, and a sequence of loss functions $\ell_t(w)$

Output: w_T

```
for t = 1, 2, ..., T do
    Observe loss ℓ_t(w_{t-1})
    Let g_t = ∇ℓ_t(w_{t-1})
    Let A_t = A_{t-1} + g_t g_t^⊤
    Let G_t = diag(A_t)^{1/2}
    Let w̃_t = w_{t-1} − η G_t^{-1} g_t
    Let w_t = arg min_{w∈Ω} (w − w̃_t)^⊤ G_t (w − w̃_t)
```

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

- Observe loss $\ell_t(w_{t-1})$
- Let $g_t = \nabla \ell_t(w_{t-1})$
- Let $A_t = A_{t-1} + g_t g_t^\top$
- Let $G_t = \mathrm{diag}(A_t)^{1/2}$
- Let $\tilde{w}_t = w_{t-1} - \eta G_t^{-1} g_t$
- Let $w_t = \arg\min_{w \in \Omega} (w - \tilde{w}_t)^\top G_t (w - \tilde{w}_t)$

Return: w_T

Regret Bound

Theorem 19 (Simplification with $p = 0.5$, Thm 15.25)

Assume that for all t, the loss function $\ell_t : \Omega \to \mathbb{R}$ is convex. Then AdaGrad method has the following regret bound:

$$\sum_{t=1}^{T} \ell_t(w_{t-1}) \leq \inf_{w \in \Omega} \sum_{t=1}^{T} \ell_t(w) + \eta \mathrm{trace}(\mathrm{diag}(A_T)^{1/2}) + \frac{\Delta_\infty^2}{2\eta} \mathrm{trace}(\mathrm{diag}(A_T)^{1/2}),$$

where $\Delta_\infty = \sup\{\|w' - w\|_\infty : w, w' \in \Omega\}$ is the L_∞-diameter of Ω.

Matrix Trace Function

The proof uses the fact that $h(B) = 2\mathrm{trace}(B^{1/2})$ is concave in B, which follows from the following result.

Theorem 20 (Thm A.18)

Let $S^d_{[a,b]}$ be the set of $d \times d$ symmetric matrices with eigenvalues in $[a, b]$. If $f(z) : [a, b] \to \mathbb{R}$ is a convex function, then

$$\mathrm{trace}(f(W))$$

is a convex function on $S^d_{[a,b]}$. This implies that for $W, W' \in S^d_{[a,b]}$:

$$\mathrm{trace}(f(W')) \geq \mathrm{trace}(f(W)) + \mathrm{trace}(f'(W)(W' - W)),$$

where $f'(z)$ is the derivative of $f(z)$.

Proof of Theorem 19 (I/II)

Consider $w \in \Omega$. The convexity of ℓ_t implies that

$$-2\eta(w_{t-1} - w)^\top g_t \le 2\eta[\ell_t(w) - \ell_t(w_{t-1})].$$

Let $G_t = \mathrm{diag}(A_t)^{1/2}$. We obtain the following result:

$$\begin{aligned}
&(\tilde{w}_t - w)^\top G_t(\tilde{w}_t - w) \\
=&(w_{t-1} - \eta G_t^{-1} g_t - w)^\top G_t(w_{t-1} - \eta G_t^{-1} g_t - w) \\
=&(w_{t-1} - w)^\top G_t(w_{t-1} - w) - 2\eta(w_{t-1} - w)^\top g_t + \eta^2 g_t^\top G_t^{-1} g_t \\
\le&(w_{t-1} - w)^\top G_t(w_{t-1} - w) + 2\eta[\ell_t(w) - \ell_t(w_{t-1})] + \eta^2 g_t^\top G_t^{-1} g_t \\
=&(w_{t-1} - w)^\top G_{t-1}(w_{t-1} - w) + (w_{t-1} - w)^\top (G_t - G_{t-1})(w_{t-1} - w) \\
&+ 2\eta[\ell_t(w) - \ell_t(w_{t-1})] + \eta^2 \mathrm{trace}((G_t^2)^{-1/2}(G_t^2 - G_{t-1}^2)) \\
\le&(w_{t-1} - w)^\top G_{t-1}(w_{t-1} - w) + (w_{t-1} - w)^\top (G_t - G_{t-1})(w_{t-1} - w) \\
&+ 2\eta[\ell_t(w) - \ell_t(w_{t-1})] + 2\eta^2[\mathrm{trace}(G_t) - \mathrm{trace}(G_{t-1})].
\end{aligned}$$

Proof of Theorem 19 (II/II)

We can use the fact that $(w_t - w)^\top G_t (w_t - w) \leq (\tilde{w}_t - w)^\top G_t (\tilde{w}_t - w)$, and then sum over $t = 1$ to $t = T$. This implies that

$$\sum_{t=1}^{T} \ell_t(w_{t-1}) \leq \sum_{t=1}^{T} \ell_t(w) + \frac{R_T}{2\eta} + \eta \left[\mathrm{trace}(G_T) - \mathrm{trace}(G_0)\right],$$

where

$$\begin{aligned} R_T =& (w_0 - w)^\top G_0 (w_0 - w) + \sum_{t=1}^{T} (w_{t-1} - w)^\top (G_t - G_{t-1})(w_{t-1} - w) \\ \leq& \Delta_\infty^2 \mathrm{trace}(G_0) + \sum_{t=1}^{T} \Delta_\infty^2 \mathrm{trace}(|G_t - G_{t-1}|) \\ \leq& \Delta_\infty^2 \mathrm{trace}(G_T). \end{aligned}$$

In the first inequality, we note that $G_t - G_{t-1}$ is a diagonal matrix.

Interpretation of Theorem 19

AdaGrad is more effective than SGD when the gradient is sufficiently sparse, which means that $\mathrm{trace}(\mathrm{diag}(A_T)^{1/2})$ can be similar to $\mathrm{trace}(\mathrm{diag}(A_T))^{1/2}$. In this case, Theorem 19 implies

$$\mathrm{trace}(\mathrm{diag}(A_T)^{1/2}) = O(\sqrt{T}).$$

Let $\eta = O(\Delta_\infty)$, then the regret bound becomes

$$O(\Delta_\infty \sqrt{T}).$$

Since in general

$$\Delta_\infty \ll \Delta_2 \equiv \sup\{\|w' - w\|_2 : w, w' \in \Omega\},$$

and in the extreme case, Δ_2 can be as large as $\Omega(\sqrt{d}\Delta_\infty)$, where d is the dimension of the model parameter. In such case, AdaGrad can be better than SGD by a factor of $\sqrt{d}$.

Summary (Chapter 15)

- Bayesian Posterior Averaging (aggregation)
- Ridge Regression (second order optimization)
- Tow approaches are closely related
- Generalization
- Aggregation Methods
- AdaGrad