

Mathematics of Machine Learning

1 Introduction

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution P_0
- classification setting $\mathcal{Y} \in \{-1, 1\}$
- regression setting $\mathcal{Y} = \mathbb{R}$

Assumption 1. $\mathcal{X} \in \mathbb{R}^p$

- hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$
- loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$
- **Classification setting**
- misclassification error $l(h(x), y) = \begin{cases} 1 & \text{if } h(x) = y \\ 0 & \text{otherwise} \end{cases}$
- classifier h
- **Regression setting**
- squared error $l(h(x), y) = (h(x) - y)^2$
- risk $R(h) = \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} l(h(x), y) dP_0(x, y)$

Fact. $R(h) = \mathbb{E}l(h(X), Y)$ for deterministic h

Setting 1. l misclassification error, R risk

- Bayes classifier h_0 — minimises misclassification risk
- Bayes risk $R(h_0)$
- regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$

Proposition 1.1. Bayes classifier h_0 , then $h_0(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$

Proof. $R(h) = \frac{1}{4}\mathbb{E}(Y - h(X))^2 = \frac{1}{4}\mathbb{E}(Y - \mathbb{E}(Y|X))^2 + \frac{1}{4}\mathbb{E}(\mathbb{E}(Y|X) - h(X))^2$ □

Setting 2. P_0 unknown

- training data (X_i, Y_i) — i.i.d. of (X, Y)
- $R(\hat{h})$

Fact. $R(\hat{h}) = \mathbb{E}(l(h(X), Y) \mid X_1, Y_1, \dots, X_n, Y_n)$

- class \mathcal{H} of hypotheses

Example. (i) $\mathcal{H} = \{h : h(x) = \text{sgn}(\mu + x^\top \beta)\}$

(ii) $\mathcal{H} = \{h : h(x) = \text{sgn}(\mu + \sum \phi_j(x)\beta_j)\}$ with dictionary $\phi_i : \mathcal{X} \rightarrow \mathbb{R}$

Setting 3. $\text{sgn}(0) = -1$

- conditional expectation $\mathbb{E}(Z \mid W)$

Proposition 1.2.

- (i) **Role of independence** $\mathbb{E}(Z|W) = \mathbb{E}Z$
- (ii) **Tower property** $\mathbb{E}[\mathbb{E}(Z|W) \mid f(W)] = \mathbb{E}[Z \mid f(W)]$
- (iii) **Taking out what is known** $\mathbb{E}(f(W)Z|W) = f(W)\mathbb{E}(Z|W)$
- (iv) **Conditional Jensen** $\mathbb{E}(f(Z)|W) \geq f(\mathbb{E}(Z|W))$ — f convex, $f(Z)$ integrable

- empirical risk / training error $\hat{R}(h) = \frac{1}{n} \sum l(h(X_i), Y_i)$
- empirical risk minimiser (ERM) $\hat{h} \in \arg \min_{h \in \mathcal{H}} \hat{R}(h)$ (multiple minimiser)
- generalisation error $R(\hat{h})$
- $h^* \in \arg \min_{h \in \mathcal{H}} R(h)$
- stochastic error / excess risk $R(\hat{h}) - R(h^*)$ — increase with complexity of \mathcal{H}
- approximation error $R(h^*) - R(h_0)$ — decrease with complexity of \mathcal{H}

Fact. $R(\hat{h}) - R(h_0) = \text{excess risk} + \text{approximation error}$

2 Statistical learning theory

Fact. $R(\hat{h}) - R(h^*) = \left(R(\hat{h}) - \hat{R}(\hat{h})\right) + \left(\hat{R}(\hat{h}) - \hat{R}(h^*)\right) + \left(\hat{R}(h^*) - R(h^*)\right)$

– concentration inequalities

Fact (Markov's inequality). W non-negative, ϕ strictly increasing, then $\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}\phi(W)}{\phi(t)}$

Fact (Chernoff bound). $\phi(t) = e^{\alpha t}$, $\alpha > 0$, then $\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t} \mathbb{E}e^{\alpha W}$

– sub-Gaussian with parameter σ — $\mathbb{E}e^{\alpha(W - \mathbb{E}W)} \leq e^{\frac{\alpha^2 \sigma^2}{2}}$

Proposition 2.1. W sub-Gaussian with σ , then $\mathbb{P}(W \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$

Proof. Chernoff bound □

Fact. W sub-Gaussian with σ , then

(i) W sub-Gaussian with σ' for all $\sigma' \geq \sigma$

(ii) $-W$ sub-Gaussian with σ

Fact. $\mathbb{P}(|W - \mathbb{E}W| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$

Proposition 2.2. W_i independent, sub-Gaussian with σ_i , ~~mean μ_i~~ , then $\gamma^\top W$ sub-Gaussian with $\sqrt{\sum_i \gamma_i \sigma_i}$

Proof. expand □

Fact. same setting, pick $\gamma = (1, \dots, 1)$, then $\mathbb{P}(\sum_i (W_i - \mu_i) \geq t) \leq \exp\left(-\frac{t^2}{2\sum_i \sigma_i^2}\right)$

Proposition 2.3. W_i mean 0, sub-Gaussian with σ (non necessarily independent), then $\mathbb{E} \max_j W_j \leq \sigma \sqrt{2 \log(d)}$

Proof. $\exp(\alpha \mathbb{E} \max W_j) \leq \mathbb{E} \exp(\alpha \max W_j) \leq \sum \exp(\alpha W_j) \leq d e^{\frac{\alpha^2 \sigma^2}{2}}$, then maximise over α □

– Rademacher r.v. ϵ — take $\{-1, 1\}$ with equal prob

Fact. Rademacher ϵ sub-Gaussian with $\sigma = 1$

Lemma 2.4 (Hoeffding's lemma). W mean 0, take values in $[a, b]$, then W sub-Gaussian with $\sigma = \frac{b-a}{2}$

Proof. weaker result $\sigma = b - a$: consider independent W' , conditional Jensen, Rademacher sub-Gaussian, $\mathbb{E}e^{\alpha W} \leq \mathbb{E}e^{\alpha \epsilon(W-W')} \leq \mathbb{E}e^{\alpha^2(W-W')^2/2} \leq \mathbb{E}e^{\alpha^2(b-a)^2/2}$ \square

– symmetrisation argument

Fact (Hoeffding's inequality). W_i independent, mean 0, $a_i \leq W_i \leq b_i$ a.s., then $\mathbb{P}(\frac{1}{n} \sum_i W_i \geq t) \leq \exp\left(-\frac{2n^2 t^2}{\sum_i (b_i - a_i)^2}\right)$

Theorem 2.5. \mathcal{H} finite, l take values in $[0, M]$, then with probability at least $1 - \delta$, $R(\hat{h}) - R(h^*) \leq M \sqrt{\frac{2(\log |\mathcal{H}| + \log \frac{1}{\delta})}{n}}$

Proof. decomposition $R(\hat{h}) - R(h^*)$, then Hoeffding's inequality \square

– $G(X_1, Y_1, \dots, X_n, Y_n) = \sup_{h \in \mathcal{H}} R(h) - \hat{R}(h)$

Fact. l takes values $[0, M]$, then $G(x_1, y_1, \dots, x_n, y_n) - G(x'_1, y'_1, x_2, y_2, \dots, x_n, y_n) \leq \frac{M}{n}$

– $a_{j:k}$ — subsequence a_j, \dots, a_k

– bound differences property:

$$f(w_1, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_n) - f(w_1, \dots, w_{i-1}, w'_i, w_{i+1}, \dots, w_n) \leq L_i$$

Theorem 2.6 (Bounded differences inequality). f bound differences property, W_i independent, then $\mathbb{P}(f(W_{1:n}) - \mathbb{E}f(W_{1:n}) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_i L_i^2}\right)$

Proof. (D_i) martingale difference wrt Doob martingale, $F_i(w_{1:i}) = \mathbb{E}(f(W_{1:n}) | W_{1:i} = w_{1:i})$
 $\begin{cases} A_i = \inf_{w_i} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}) | W_{1:i-1}) \\ B_i = \sup_{w_i} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}) | W_{1:i-1}) \end{cases}$, then use $W_{(i+1:n)}$ independent to W_i , then Azuma-Hoeffding \square

– martingale sequence $(Z_i)_{i \geq 0}$ wrt $(W_i)_{i \geq 0}$ —

(i) $\mathbb{E}|Z_i| < \infty$

(ii) Z_i $\sigma(W_{0:i})$ -measurable

(iii) $\mathbb{E}(Z_i | W_{0:(i-1)}) = Z_{i-1}$

– martingale difference sequence $D_i = Z_i - Z_{i-1}$

– Doob martingale $Z_i = \mathbb{E}f(W_{1:n}) | W_{1:i}$ — martingale provided $\mathbb{E}|f(W_{1:n})| < \infty$

Lemma 2.7. (D_i) martingale difference sequence wrt (W_i) , $\mathbb{E}(e^{\alpha D_i} | W_{0:i-1}) \leq e^{\frac{\alpha^2 \sigma_i^2}{2}}$,
then $\gamma^\top D$ sub-Gaussian with $\sqrt{\sum \gamma_i^2 \sigma_i^2}$

Proof. Tower property with $\sigma(W_{1:i})$ for $i = n-1, n-2, \dots, 1$ \square

Theorem 2.8 (Azuma-Hoeffding). (D_i) martingale difference sequence wrt (W_i) ,
 $\exists \sigma(W_{0:(i-1)})$ -measurable A_i, B_i , constant L_i st

(i) $A_i \leq D_i \leq B_i$ a.s.

(ii) $B_i - A_i \leq L_i$

, then $\mathbb{P}(\sum_i D_i \geq t) \leq \exp\left(-\frac{2t^2}{\sum L_i^2}\right)$

Proof. Hoeffding's Lemma conditionally on $W_{0:(i-1)}$, then lemma, then Gaussian tail bound \square

Setting 4. \mathcal{H} (possibly infinite) hypothesis class, l takes values in $[0, M]$

Fact. $R(\hat{h}) - R(h^*) \leq (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*)$

– $Z_i = (X_i, Y_i)$

– $\mathcal{F} = \{(x, y) \mapsto -l(h(x), y) : h \in \mathcal{H}\}$

Fact. $G = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum (f(Z_i) - \mathbb{E}f(Z_i))$

– $\mathcal{R}_n(\mathcal{F}) = \mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \epsilon_i f(Z_i) \right)$ — ϵ_i i.i.d. Rademacher independent of $Z_{1:n}$

Intuition. capture how closely $f(Z_i)$ align with random label ϵ_i (dot product)

Theorem 2.9. \mathcal{F} class of real functions, Z_i i.i.d. ,
then $\mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum (f(Z_i) - \mathbb{E}f(Z_i)) \right) \leq 2\mathcal{R}_n(\mathcal{F})$

Proof. Z'_i i.i.d. copy of Z_i , symmetrisation technique:
 $\sup \frac{1}{n} \sum f(Z_i) - \mathbb{E}f(Z_i) \leq \mathbb{E} \left(\sup \frac{1}{n} \sum f(Z_i) - f(Z'_i) | Z_{1:n} \right)$ \square

– $\mathcal{F}(z_{1:n}) = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\}$

– empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \epsilon_i f(z_i) \right)$

– $\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})) = \mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \epsilon_i f(Z_i) \mid Z_{1:n} \right)$

Fact. $\mathcal{R}_n(\mathcal{F}) = \mathbb{E} \hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))$

Theorem 2.10 (Generalisation bound based on Rademacher complexity).

$\mathcal{F} = \{(x, y) \mapsto l(h(x), y)\}$, l takes values in $[0, M]$,

then with probability at least $1 - \delta$, $R(\hat{h}) - R(h^*) \leq 2\mathcal{R}_n(\mathcal{F}) + M\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$

Proof. decomposition: $R(\hat{h}) - R(h^*) \leq (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*)$

Bounded differences inequality: $\mathbb{P}(G - \mathbb{E}G \geq \frac{t}{2}) \leq \exp\left(-\frac{t^2 n}{2M^2}\right)$,

Hoeffding's inequality: $\mathbb{P}\left(\hat{R}(h^*) - R(h^*) \geq \frac{t}{2}\right) \leq \exp\left(-\frac{t^2 n}{2M^2}\right)$

$\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, so $\mathbb{E}G \leq 2\mathcal{R}_n(\mathcal{F})$, then $t = M\sqrt{\frac{2\log \frac{1}{\delta}}{n}}$ □

Setting 5. classification setting, misclassification loss, $\mathcal{F} = \{(x, y) \mapsto l(h(x), y) : h \in \mathcal{H}\}$

Fact. $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(x_{1:n})|$

Lemma 2.11. $\hat{R}(\mathcal{F}(z_{1:n})) \leq \sqrt{\frac{2\log |\mathcal{F}(z_{1:n})|}{n}} = \sqrt{\frac{2\log |\mathcal{H}(x_{1:n})|}{n}}$

Proof. $\mathcal{F}' = \{f_1, \dots, f_d\}$ st $\mathcal{F}'(z_{1:n}) = \mathcal{F}(z_{1:n})$, $W_j = \frac{1}{n} \sum \epsilon_i f_j(z_i)$, then W_j sub-Gaussian with $\sigma = \frac{1}{\sqrt{n}}$, then apply max bound □

Setting 6. \mathcal{F} class of functions $f : \mathcal{X} \mapsto \{a, b\}$, $\mathcal{F} \geq 2$

- \mathcal{F} shatters $x_{1:n}$ — $|\mathcal{F}(x_{1:n})| = 2^n$
- shattering coefficient $s(\mathcal{F}, n) = \max_{x_{1:n}} |\mathcal{F}(x_{1:n})|$
- VC dimension $VC(\mathcal{F}) = \sup \{n : s(\mathcal{F}, n) = 2^n\}$

Lemma 2.12 (Sauer-Shelah). $VC(\mathcal{F}) = d$, then $s(\mathcal{F}, n) \leq \sum_0^d \binom{n}{i} \leq (n+1)^d$

Proof. non-empty $Q \subset [n]$, stronger statement: at least $|\mathcal{F}(x_{1:n})| - 1$ non-empty Q st \mathcal{F} shatters x_Q , then induction on $|\mathcal{F}(x_{1:n})|$ □

Fact. $\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2VC(\mathcal{F})\log(n+1)}{n}}$

Setting 7. \mathcal{F} vector space of functions, $\mathcal{H} = \{h : h(x) = \text{sgn}(f(x)), f \in \mathcal{F}\}$

Example. $\mathcal{X} = \mathbb{R}^p$, $\mathcal{F} = \{x \mapsto x^\top \beta : \beta \in \mathbb{R}^p\}$

Proposition 2.13. Under above setting, $VC(\mathcal{H}) \leq \dim(\mathcal{F})$

Proof. $d = \dim(\mathcal{F}) + 1$, linear map $L(f) = (f(x_1), \dots, f(x_d))$, then $\sum_{\gamma_i > 0} \gamma_i f(x_i) + \sum_{\gamma_i} f(x_i) = 0$, then pick h forcing contradiction, so $x_{1:d}$ cannot be shattered □

3 Computation for empirical risk minimisation

- convex set C — $x, y \in C$, then $(1-t)x + ty \in C$ for all $t \in (0, 1)$
- convex function f — $f : C \rightarrow \mathbb{R}$, $f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$ for all $x, y \in C, t \in (0, 1)$
- concave function — $-f$
- strictly convex

Fact (Local to global phenomenon). *local minimum \Rightarrow global minimum*

- Hessian matrix at x $H(x)$

Proposition 3.1. *C convex set, f convex function, then*

- (i) *g convex, $a, b \geq 0$, then $af + bg$ convex function*
- (ii) *A matrix, b vector, $C = \mathbb{R}^d$, then $g(x) = f(Ax - b)$ convex function*
- (iii) *I index set, f_α convex for $\alpha \in I$, $g(x) = \sup_{\alpha \in I} f_\alpha(x)$, then*
 - (a) *$D = \{x : g(x) < \infty\}$ convex*
 - (b) *g restricted to D convex*
- (iv) *f differentiable at $x \in \text{int}(C)$, then $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$*
- (v) *f twice differentiable, then*
 - (a) *f convex $\iff H(x)$ positive semi-definite*
 - (b) *f strictly convex $\iff H(x)$ positive definite*

Setting 8. Classification framework:

- (i) *family \mathcal{H} of h*
- (ii) *each h determine classifier by $x \mapsto \text{sgn}(h(x))$*
- (iii) *loss function $l(h(x), y) = \phi(yh(x))$ where ϕ convex and aim to approximate $\mathbb{1}_{(\infty, 0]}$*
- (iv) *ϕ -risk $R_\phi = \mathbb{E}(\phi(Yh(X)))$*

Example (Surrogate loss).

- (i) **Hinge loss:** $\phi(u) = \max(1 - u, 0)$
- (ii) **Exponential loss:** $\phi(u) = e^{-u}$
- (iii) **Logistic loss:** $\phi(u) = \log_2(1 + e^{-u})$

- $h_{\phi,0}$ ERM of surrogate loss
- $\eta(x) = \mathbb{P}(Y = 1|X = x)$

Idea. want $x \mapsto \text{sgn}(h_{\phi,0}(x))$ mimics Bayes classifier $x \mapsto \text{sgn}(\eta(x) - \frac{1}{2})$

- conditional ϕ -risk $\mathbb{E}(\phi(Yh(X))|X=x) = \eta(x)\phi(h(x)) + (1-\eta(x))\phi(-h(x))$
- $C_\eta(\alpha) = \eta(x)\phi(\alpha) + (1-\eta(x))\phi(-\alpha) = \mathbb{E}(\phi(Y\alpha))$
- classification calibrated — $\inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) < \inf_{\alpha(2\eta-1) \leq 0} C_\eta(\alpha)$ for all $\eta \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$

Theorem 3.2. ϕ convex, then ϕ classification calibrated \iff differentiable at 0, $\phi'(0) < 0$

Proof. $C'_\eta(0) = (2\eta - 1)\phi'(0)$, assume $\eta > \frac{1}{2}$, $C_\eta(\alpha) \geq C_\eta(0)$ for $\alpha \leq 0$, then find $\alpha^* > 0$ st $C_\eta(\alpha^*) < C_\eta(0)$ \square

Setting 9. $\mathcal{F} = \{(x, y) \mapsto \phi(yh(x)) : h \in \mathcal{H}\}$

Lemma 3.3 (Contraction lemma). $r = \sup_{x \in \mathcal{X}, h \in \mathcal{H}} |h(x)|$, $\exists L \geq 0$, $|\phi(u) - \phi(u')| \leq L|u - u'|$ for $u, u' \in [-r, r]$ (Lipschitz with L on $[-r, r]$), then $\mathcal{R}_n(\mathcal{F}) \leq L\mathcal{R}_n(\mathcal{H})$

Proof. $\mathbb{E} \sup_h (\frac{1}{n} \epsilon_i \phi(y_i h(x_i)) + A(h, \epsilon_{-i})) \leq \mathbb{E} \sup_h (\frac{L}{n} \epsilon_i h(x_i) + A(h, \epsilon_{-i}))$, then stepwise argument inequality from conditioning ϵ_{-i} and expand ϵ_i \square

Corollary 3.4. setup of contraction lemma, r finite, ϕ non-increasing, $M = \phi(-r)$, then with probability at least $1 - \delta$, $R_\phi(\hat{h}) - R_\phi(h^*) \leq 2L\mathcal{R}_n(\mathcal{H}) + M\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$

Example (l_2 -constraint).

Setting 10. $\mathcal{X} = \{\|x\|_2 \leq C\}$, $\mathcal{H} = \{x \mapsto x^\top \beta : \|\beta\|_2 \leq \lambda\}$

Fact. $\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \frac{\lambda C}{\sqrt{n}}$ (Cauchy-Schwarz, Jensen)

Fact. $\sup_{x,h} |h(x)| = \lambda C$

Example (l_1 -constraint).

Setting 11. $\mathcal{X} = \{\|x\|_\infty \leq C\}$, $\mathcal{H} = \{x \mapsto x^\top \beta : \|\beta\|_1 \leq \lambda\}$

- convex hull $\text{conv } S$ — intersection of all convex sets containing S

- convex combination $v = \sum \alpha_i v_i$ — $\sum \alpha_i = 1$

Lemma 3.5. $v \in \text{conv } S \iff v$ convex combination of points in S

Proof. induction \square

Lemma 3.6. L linear map, then $\text{conv } L(S) = L(\text{conv } S)$

Lemma 3.7. $\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}(\text{conv } A)$

- $S = \bigcup_{j=1}^p \{\lambda e_j, -\lambda e_j\}$
- $L(\beta) = (x_1^\top \beta, \dots, x_n^\top \beta)^\top$

Fact. $\{\|\beta\|_1 \leq \lambda\} = \text{conv } S$

Fact. $\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) = \hat{\mathcal{R}}(L(S)) = \frac{\lambda}{n} \mathbb{E}(\max |\sum \epsilon_i x_{ij}|) \leq \frac{\lambda C}{\sqrt{n}} \sqrt{2 \log(2p)}$ (sub-Gaussian bound for max)

Fact. $\sup |h(x)| = \lambda C$

Proposition 3.8. C closed convex set, then

- (i) minimiser of $\|x - z\|_2$ exists and unique
- (ii) let $\pi_C(x) = \arg \min_{z \in C} \|x - z\|_2$, then
 - $(x - \pi_C(x))^\top (z - \pi_C(x)) \leq 0$ for all $z \in C$
 - $\|\pi_C(x) - \pi_C(y)\|_2 \leq \|x - y\|_2$ for all $y \in \mathbb{R}^d$

Proof. (i) **Existence:** bounded set $B = \{w : \|w - x\|_2 \leq \inf \|x - z\|_2 + 1\}$

(ii) **Uniqueness:** $z \mapsto \|x - z\|_2^2$ convex

□

- projection $\pi_C(x)$

Proposition 3.9. C closed convex set, $x \notin C$, then $\exists v, \epsilon > 0$ st $v^\top z \leq v^\top x - \epsilon$

Proof. $v = x - \pi_C(x)$

□

- subgradient g — f convex, $f(z) \geq f(x) + g^\top(z - x)$
- subdifferential $\partial f(x)$ — set of subgradients
- epigraph $C = \{(z, y) : y \geq f(z)\}$

Proposition 3.10. f convex, then $\partial f(x)$ non-empty for all x

Proof. epigraph closed, convex, $w_k \notin C \rightarrow (x, f(x))$, apply prop, BW □

Proposition 3.11. f convex, f differentiable at x , then $\partial f(x) = \{\nabla f(x)\}$

Proof. g subgradient, then $\lim_{t \rightarrow 0} \frac{f(x+tz) - f(x)}{t} \geq g^\top z$ □

Proposition 3.12 (Subgradient calculus). f, f_1, f_2 convex, $h(x) = f(Ax + b)$, then

(i) $\partial(\alpha f)(x) = \{\alpha g : g \in \partial f(x)\}$

(ii) $\partial(f_1 + f_2)(x) = \{g_1, g_2 : g_i \in \partial f_i(x)\}$

(iii) $\partial h(x) = A^\top \partial f(Ax + b)$

Gradient descent

– Parameters:

- $\beta_1 \in C$, k , step sizes (η_s)

– Procedures:

For $s = 1, \dots, k - 1$:

(i) compute $g_s \in \partial f(\beta_s)$

(ii) $z_{s+1} = \beta_s - \eta_s g_s$

(iii) $\beta_{s+1} = \pi_C(z_{s+1})$

– Return:

- $\bar{\beta} = \frac{1}{k} \sum \beta_s$

Theorem 3.13. f convex function, C closed convex, $\hat{\beta}$ minimiser of f over C , $\sup_{\beta \in C} \|\beta\|_2 \leq R$, $\sup_{\beta \in C} \sup_{g \in \partial f(\beta)} \|g\|_2 \leq L$, step size $\eta_s = \eta = \frac{2R}{L\sqrt{k}}$, then $f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{2LR}{\sqrt{k}}$

Proof. Jensen □

Stochastic gradient descent

Setting 12. $f(\beta) = \mathbb{E} \tilde{f}(\beta; U)$, $\beta \mapsto \tilde{f}(\beta; u)$ convex for all u

– Parameters:

- $\beta_1 \in C$, k , step sizes (η_s) , U_i i.i.d.

– Procedures:

For $s = 1, \dots, k - 1$:

- (i) compute $\tilde{g}_s \in \partial \tilde{f}(\beta_s; U_s)$
- (ii) $z_{s+1} = \beta_s - \eta_s \tilde{g}_s$
- (iii) $\beta_{s+1} = \pi_C(z_{s+1})$

Theorem 3.14. f convex function, C closed convex, $\hat{\beta}$ minimiser of f over C , $\sup_{\beta \in C} \|\beta\|_2 \leq R$, $\sup_{\beta \in C} \mathbb{E} \left(\sup_{\tilde{g} \in \partial \tilde{f}(\beta; U)} \|\tilde{g}\|_2^2 \right) \leq L^2$, step size $\eta_s = \eta = \frac{2R}{L\sqrt{k}}$, then $\mathbb{E}f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{2LR}{\sqrt{k}}$

Proof. $g_s = \mathbb{E}(\tilde{g}_s \mid \beta_s) \in \partial f(\beta_s)$ □

4 Popular machine learning methods

- cross validation — minimise $CV(j) = \frac{1}{n} \sum_{k=1}^v \sum_{i \in A_k} l(H_{-k}^j(X_i), Y_i)$ over j
- stacking — minimise $\frac{1}{n} \sum_{k=1}^v \sum_{i \in A_k} l(\sum_{j=1}^m w_j H_{-k}^j(X_i), Y_i)$ over w on $\{w : w_j \geq 0\}$
- Adaboost

Setting 13. $h : \mathcal{X} \rightarrow \{-1, 1\}$, $\mathcal{H}^M = \left\{ \sum_{m=1}^M \beta_m h_m \right\}$

- tuning parameters M

Procedures:

- (i) \hat{f}_0 — $x \mapsto 0$
- (ii) $(\hat{\beta}_m, \hat{h}_m) = \arg \min \frac{1}{n} \sum \exp \left[-Y_i (\hat{f}_{m-1}(X_i) + \beta h(X_i)) \right]$
- (iii) $\hat{f}_m = \hat{f}_{m-1} + \hat{\beta}_m \hat{h}_m$

Return:

- $\text{sgn} \circ \hat{f}_M$
- $err_m(h) = \frac{\sum w_i^{(m)} \mathbb{1}_{h(X_i) \neq Y_i}}{\sum w_i^{(m)}}$

Fact. $\hat{h}_m = \arg \min err_m(h)$, $\hat{\beta}_m = \frac{1}{2} \log \left(\frac{1 - err_m(\hat{h}_m)}{err_m(\hat{h}_m)} \right)$

Example. – decision stumps $\mathcal{H} = \{h_{a,j,1}(x) = \text{sgn}(x_j - a), h_{a,j,2}(x) = \text{sgn}(a - x_j)\}$