Algorithmic Foundations of Learning

1 Offline Statistical Learning: Prediction

The standard framework:

1. observe training examples $S = \{Z_1, ..., Z_n\} \in \mathcal{Z}^n$, assumed iid, unknown dist
2. apply some decision function to S (and possibly some source of randomness) to choose $A \in \mathcal{A} \subseteq \mathcal{B}$, where $\mathcal{B}$ is the set of all actions, and $\mathcal{A}$ the set of admissible actions (i.e. we have limited ourselves)
3. minimise & control estimation error as a function of n, "complexity" of $\mathcal{A}$, where:
   - prediction loss function is $\ell : \mathcal{B} \times \mathcal{Z} \to \mathbb{R}_+$
   - $r : \mathcal{B} \to \mathbb{R}_+$ is the expected/population risk, def by $r(a) := \mathbb{E}[\ell(a, Z)]$, where $Z \in \mathcal{Z}$ is a new, independent test data point (from the same dist)
   - **excess risk** is $r(A) - \inf_{a\in\mathcal{B}} r(a) = \underbrace{r(A) - \inf_{a\in\mathcal{A}} r(a)}_{\textbf{estimation error}} + \underbrace{\inf_{a\in\mathcal{A}} r(a) - \inf_{a\in\mathcal{B}} r(a)}_{\text{approximation error}}$

notes on this:

- if A is a random action (not a chosen one), then $r(A) \neq \mathbb{E}[\ell(A, Z)]$, but $r(A) = \mathbb{E}[\ell(A, Z)|A]$,so that it is actually a random variable
- we assume minimisers of the $\inf$'s exist, precisely $a^\star \in \arg\min_{a\in\mathcal{A}} r(a), a^{\star\star} \in \arg\min_{a\in\mathcal{B}} r(a)$
- irreducible risk: $r(a^{\star\star})$, as we cannot reduce past this.
- **bias-complexity tradeoff = approximation-estimation tradeoff**, as balancing the two errors in the excess risk
- $\mathcal{A}$ is a reflection of our prior knowledge of the system, and is useful to:
  - implement regularisation to avoid overfitting, as S will only contain a finite number of samples
  - **No Free Lunch theorem** - don't understand point!

Supervised learning

So $Z_i := (X_i, Y_i)$, where X_i is the **feature/covariate**, and Y_i the **label**.

$\mathcal{B} := \{a : \mathcal{X} \to \mathcal{Y}\}$, so the loss function must be of the form $\ell(a, x, y) := \phi(a(x), y)$, and we call the chosen $a \in \mathcal{A} \subseteq \mathcal{B}$ the **predictor/hypothesis**.

we call $a^{\star\star} = \arg\min_{a\in\mathcal{B}} \mathbb{E}[\phi(a(X), Y)]$ the **Bayes decision rule**, and $r(a^{\star\star}) = \mathbb{E}[\phi(a^{\star\star}(X), Y]$ is the **Bayes risk**. There is a deterministic solution to these (though it is not computable, as we do not know the distribution of $Y|X$), given by:

$$a^{\star\star}(x) = \arg\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}[\phi(\hat{y}, Y)|X = x]$$

(prove by showing $r(a^{\star\star}) \leq r(a)$ for all a

Regression

Given $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}$, we have regression. With the ℓ_2 loss function, $\phi(\hat{y}, y) = (\hat{y} - y)^2$ and $a^{\star\star}(x) = \mathbb{E}[Y|X = x]$ *see sheet 1*, and with the ℓ_1 loss function, $\phi(\hat{y}, y) = |\hat{y} - y|$ and $a^{\star\star}(x) = \text{Median}[Y|X = x]$.

We can choose $\mathcal{A}$ to definethe solution type, whtether it is linear, neural/kernel methods/etc.

In this case, the expected excess risk also decomposes to the **bias-variance tradeoff**:

$$\mathbb{E}[r(A) - r(a^{\star\star})] = \underbrace{\mathbb{E}\left[(\mathbb{E}[A(X)|X] - a^{\star\star}(X))^2\right]}_{\text{expected squared bias}} + \underbrace{\mathbb{E}\left[\text{Var}[A(X)|X]\right]}_{\text{expected variance}}$$

- the expected variance is a measure of how much $A(X)$ changes if X changes (which increases as $\mathcal{A}$ gets "larger")
- the expected squared bias is a measure of the error on X given by choosing A instead of the best model $a^{\star\star} \in \mathcal{B}$

Classification

A subset of regression, where now $\mathcal{Y} = \{y_1, ..., y_k\}$, representing labels. There are many possible loss functions:

- **zero-one loss:** $\phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$ gives the Bayes decision rule = MAP = $\arg\max_{\hat{y} \in \mathcal{Y}} \mathbb{P}(Y = \hat{y}|X = x)$
- **exponential** loss: $\phi(u) = e^{-u}$, specifically for the case $\mathcal{Y} = \{-1, 1\}$, where we write $\phi(\hat{y}, y) = \phi(\hat{y} \cdot y)$
- **hinge loss** $\phi(u) = \max\{1 - u, 0\}$
- **logistic loss** $\phi(u) = \log_2(1 + e^{-u})$
- these last three all bound the zero-one loss above (which is $\mathbf{1}_{u \leq 0}$ in this case)

ERM

Empirical Risk Minimization uses the **empirical risk function** $R : \mathcal{B} \to \mathbb{R}_+$ $R(a) := \frac{1}{n}\sum_{i=1}^n \ell(a, Z_i)$, which converges a.s. to $r(a)$ as $n \to \infty$ by the Law of Large Numbers

$A^\star \in \arg\min_{a \in \mathcal{A}} R(a)$ is one of the minimisers of the ERF (and so is a RV because it is based on the data samples)

We want bounds on the estimation error $r(A^\star) - r(a^\star)$ when we choose A as $A^\star$, in two forms:

- $\mathbb{E}[r(A^\star) - r(a^\star)] \leq$ ExBound, a positive quantity that depends on $n, \mathcal{A}$
- $\mathbb{P}(r(A^\star) - r(a^\star) \geq \varepsilon) \leq \text{UpperTail}(\varepsilon)$, a strictly decr. function of ε
  - or equiv., $\mathbb{P}(r(A^\star) - r(a^\star) < \text{UpperTail}^{-1}(\delta)) \geq 1 - \delta = 1 - \text{UpperTail}(\varepsilon)$

Uniform learning decomposition:

$$r(A^\star) - r(a^\star) \leq \underbrace{R(A) - R(A^\star)}_{\leq 0} + \sup_{a \in \mathcal{A}}\{r(a) - R(a)\} + \sup_{a \in \mathcal{A}}\{R(a) - r(a)\}$$

Approximations to ERM minimizers

let A be some approximation (RV) to A^*, when A^* is intractable to compute

$$\begin{aligned} r(A) - r(a^*) &= (r(A) - R(A)) + (R(A) - R(A^*)) + \underbrace{(R(A^*) - R(a^*))}_{\leq 0} + (R(a^*) - r(a^*)) \\ &\leq \underbrace{(R(A) - R(A^*))}_{\text{Optimisation}} + \underbrace{\sup_{a \in \mathcal{A}}\{r(a) - R(a)\} + \sup_{a \in \mathcal{A}}\{r(a) - R(a)\}}_{\text{Statistics}} \end{aligned}$$

And each of these terms can now be bounded in expectation or with tail probability bounds.

TODO - PAGES 1-9,1-10

Other problems

Estimation: given a data sample from D with parameter $a^* \in \mathcal{A}$, estimate $A \in \mathcal{A}$, and suffer a loss $\ell(A, a^*)$ - tradeoff: complexity of $\mathcal{A}$ vs estimation error in terms of n

Online learning: same as CLT. **fill in later......**

2 Maximal inequalities and Rademacher complexity

Synopsis: some bounds, def of Rademacher complexity, how to use it to bound error decompositions

Lemma 2.1 (Hoeffding's lemma). *X a real-valued RV st $a \leq X - \mathbb{E}X \leq b$. $\forall \lambda \in \mathbb{R}$*

$$\mathbb{E} \exp \lambda(X - \mathbb{E}X) \leq \exp(\lambda^2(b-a)/8)$$

Lemma 2.2. *Application of Hoeffding's lemma to n variables: X_i as in Hoeffding's:*

$$\mathbb{E} \max_{i \in [n]} X_i - \mathbb{E}X_i \leq \frac{b-a}{2}\sqrt{2 \log n}$$

Proposition 2.3. *$\ell(a, z) \in [0, 1]$ for all a, z*

$$\mathbb{E} \max_{a \in \mathcal{A}} (r(a) - R(a)) \leq \sqrt{\frac{2 \log |\mathcal{A}|}{n}}$$

Definition 2.4 (Rademacher Complexity).

$$\operatorname{Rad}(\mathcal{T}) := \mathbb{E}_{\Omega_i} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i t_i$$

(nb this is equiv to empirical R-C in CLT, as in CLT we don't define it over sets)

Proposition 2.5 (Properties of Rademacher complexity).

$$\operatorname{Rad}(c\mathcal{T} + d) = |c| \operatorname{Rad}(\mathcal{T})$$

$$\operatorname{Rad}(\mathcal{T} + \mathcal{T}') = \operatorname{Rad}(\mathcal{T}) + \operatorname{Rad}(\mathcal{T}')$$

$$\operatorname{Rad}(\operatorname{conv}(\mathcal{T})) = \operatorname{Rad}(\mathcal{T})$$

$\operatorname{conv}(\mathcal{T})$:=convex hull, usual definition.

Lemma 2.6 (Massart's Lemma). *for $\mathcal{T} \in \mathbb{R}^n$, fixed $t_0 \in \mathcal{T}$.*

$$\operatorname{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - t_0\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n}$$

Lemma 2.7 (Talagrand's Lemma). *if f is γ-Lipschitz (or f_i are, 1 per coord)*

$$\operatorname{Rad}(f \circ \mathcal{T}) \leq \gamma \operatorname{Rad}(\mathcal{T})$$

Proposition 2.8 (Symmetrization).

$$\mathbb{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq 2\mathbb{E} \operatorname{Rad}(\mathcal{L} \circ \{Z_1, ..., Z_n\})$$

where $\mathcal{L} \circ S := \{(\ell(a, z_i), ..., \ell(a, z_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$, as we define the class of functions $\mathcal{L} := \{Z \ni z \mapsto \ell(a, z) \in \mathbb{R} : a \in \mathcal{A}\}$

Definition 2.9 (Rademacher complexity for classes of functions)**.** Empirical Rademacher complexity of $\mathcal{L}$ at a (random) sample $Z_1, ..., Z_n$

$$\mathrm{Rad}(\mathcal{L} \circ \{Z_1, ..., Z_n\}) = \mathbb{E}\left[\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n} \Omega_i \ell(a, Z_i)\middle| Z_1, ..., Z_n\right]$$

Rademacher complexity of $\mathcal{L}$:

$$\mathrm{Rad}(\mathcal{L}) := \mathbb{E}\left[\mathrm{Rad}(\mathcal{L} \circ \{Z_1, ..., Z_n\})\right]$$

3 More Rademacher complexity

Synopsis: practical bounds on $\mathrm{Rad}(\mathcal{L} \circ \{Z_1, ..., Z_n\})$ for different problem setups

Proposition 3.1. *$\hat{y} \mapsto \phi(\hat{y}, y)$ is γ-Lipschitz for any $y \in \mathcal{Y}$: forall samples (x_i, y_i):*

$$\mathrm{Rad}\left(\mathcal{L} \circ \{(x_1, y_1), ..., (x_n, y_n)\}\right) \le \gamma\,\mathrm{Rad}(\mathcal{A} \circ \{x_1, ..., x_n\})$$

Proposition 3.2 (ℓ_2, ℓ_2)**.** $\mathcal{A}_2 := \left\{x \in \mathbb{R}^d \mapsto w^\top x : \|w\|_\mathbf{2} \le c\right\}$

$$\mathrm{Rad}(\mathcal{A}_2 \circ \{x_1, ..., x_n\}) \le c \max_i \|x_i\|_2 \frac{1}{\sqrt{n}} \le c \max_i \|x_i\|_\infty \frac{\sqrt{d}}{\sqrt{n}}$$

Proposition 3.3 (ℓ_1, ℓ_∞)**.** $\mathcal{A}_1 := \left\{x \in \mathbb{R}^d \mapsto w^\top x : \|w\|_1 \le c\right\}$

$$\mathrm{Rad}(\mathcal{A}_1 \circ \{x_1, ..., x_n\}) \le c \max_i \|x_i\|_\infty \frac{\sqrt{2\log(2d)}}{\sqrt{n}}$$

Proposition 3.4 (simplex, ℓ_∞)**.** $\mathcal{A}_\Delta := \left\{x \in \mathbb{R}^d \mapsto w^\top x : \|w\|_1 \le 1, w_i \ge 0\right\}$

$$\mathrm{Rad}(\mathcal{A}_\Delta \circ \{x_1, ..., x_n\}) \le \max_i \|x_i\|_\infty \frac{\sqrt{2\log d}}{\sqrt{n}}$$

TODO: same for neural networks

4 VC dimension

Definition 4.1 (Growth function)**.** of $\mathcal{A} \subseteq \{a : X \to \pm 1\}$:

$$\tau_\mathcal{A}(n) := \sup_{x\in\mathcal{X}^n} |\mathcal{A} \circ x| \le 2^n$$

Proposition 4.2.

$$\mathrm{Rad}(\mathcal{A} \circ \{x_1, ..., x_n\} \le \sqrt{\frac{2\log\tau_\mathcal{A}(n)}{n}}$$

Definition 4.3 (VC dimension)**.**

$$\mathrm{VCD}(\mathcal{A}) := \max\{n \in \mathbb{N} : \tau_\mathcal{A}(n) = 2^n\}$$

Lemma 4.4 (Sauer-Shelah's)**.**

$$\tau_\mathcal{A}(n) \le \frac{en}{\mathrm{VCD}(\mathcal{A})}^{\mathrm{VCD}(\mathcal{A})}$$

Proposition 4.5. *for any $\{x_1, ..., x_n\}$:*

$$\mathrm{Rad}(\mathcal{A} \circ x) \le \sqrt{\frac{2\,\mathrm{VCD}(\mathcal{A})\log(en/\,\mathrm{VCD}(\mathcal{A})}{n}}$$

Can use this for boundind $\mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \le \mathbb{E}\,\mathrm{Rad}(\mathcal{A} \circ \{X_1, ..., X_n\})$, as bound independent of data.

5 Covering and packing numbers (CH4.5,5-5.??)

pseudo-metric - a metric except $\rho(x, y) = 0 \not\Longrightarrow x = y$

Definition 5.1 (Covering and packing numbers)**.** given $(\mathcal{S}, \rho)$ is a pseudo-metric space, and $\varepsilon > 0$:

$\mathcal{C} \subseteq \mathcal{S}$ is an ε-cover of $(\mathcal{S}, \rho)$ if $\forall x \in \mathcal{S}\ \exists y \in \mathcal{C}$ st $\rho(x, y) \leq \varepsilon$. $\mathcal{C} \subseteq \mathcal{S}$ is a minimal ε-cover if it has the smallest cardinality, and $\mathrm{Cov}(\mathcal{S}, \rho, \varepsilon)$ is the size of any minimal ε-cover.

$\mathcal{P} \subseteq \mathcal{S}$ is an ε-packing of $(\mathcal{S}, \rho)$ if $\forall x, x' \in \mathcal{P}\ \rho(x, x') > \varepsilon$. $\mathcal{P} \subseteq \mathcal{S}$ is a maximal ε-cover if it has the largest cardinality, and $\mathrm{Pack}(\mathcal{S}, \rho, \varepsilon)$ is the size of any maximal ε-cover.

Proposition 5.2 (Duality between covering and packing)**.**

$$\mathrm{Cov}(\mathcal{S}, \rho, \varepsilon) \leq \mathrm{Pack}(\mathcal{S}, \rho, \varepsilon) \leq \mathrm{Cov}(\mathcal{S}, \rho, \varepsilon/2)$$

Proposition 5.3 (Bounded balls: coverings and packings)**.** *Considering the normed space* $(\mathbb{R}^d, \|\cdot\|)$ *for a norm* $\|\cdot\|$*, if* $\varepsilon \leq r$*:*

$$\left(\frac{r}{\varepsilon}\right)^d \leq \mathrm{Cov}(\mathcal{B}(x, r), \|\cdot\|, \varepsilon) \leq \mathrm{Pack}(\mathcal{B}(x, r), \|\cdot\|, \varepsilon) \leq \left(\frac{3r}{\varepsilon}\right)^d$$

Definition 5.4 (Data-dependent pseudonorms)**.** Given data points $\boldsymbol{x} := \{x_1, ..., x_n\} \in \mathcal{X}^n$, define the following pseudo-norms, which induce pseudo-metrics

$$\|a\|_{p,\boldsymbol{x}} := \left(\frac{1}{n}\sum_{i=1}^{n} |a(x_i)|^p\right)$$
$$\|a\|_{\infty,\boldsymbol{x}} := \max_{i\in[n]} |a(x_i)$$

Proposition 5.5 (monotonicity of covering and packing numbers)**.** *for any* $\boldsymbol{x} := \{x_1, ..., x_n\} \in \mathcal{X}^n, 1 \leq p \leq q$ *and* $\varepsilon \geq 0$*:*

$$\mathrm{Cov}(\mathcal{A}, \|\cdot\|_{p,\boldsymbol{x}}, \varepsilon) \leq \mathrm{Cov}(\mathcal{A}, \|\cdot\|_{q,\boldsymbol{x}}, \varepsilon)$$
$$\mathrm{Pack}(\mathcal{A}, \|\cdot\|_{p,\boldsymbol{x}}, \varepsilon) \leq \mathrm{Pack}(\mathcal{A}, \|\cdot\|_{p,\boldsymbol{x}}, \varepsilon)$$

Proposition 5.6 (Rademacher complexity and Covering numbers)**.** *for* $\boldsymbol{x} := \{x_1, ..., x_n\} \in \mathcal{X}^n$, $c_{\boldsymbol{x}}$ *st* $\sup_{a\in\mathcal{A}} \|a\|_{2,\boldsymbol{x}} \leq c_{\boldsymbol{x}}$*:*

$$\mathrm{Rad}(\mathcal{A} \circ \boldsymbol{x}) \leq \inf_{\varepsilon>0} \left\{\varepsilon + \frac{\sqrt{2}c_{\boldsymbol{x}}}{\sqrt{n}}\sqrt{\log \mathrm{Cov}(\mathcal{A}, \|\cdot\|_{1,\boldsymbol{x}}, \varepsilon)}\right\}$$

Proposition 5.7 (Chaining: Rademacher & Covering numbers again)**.** *for* $\boldsymbol{x} := \{x_1, ..., x_n\} \in \mathcal{X}^n$, $c_{\boldsymbol{x}}$ *st* $\sup_{a\in\mathcal{A}} \|a\|_{2,\boldsymbol{x}} \leq c_{\boldsymbol{x}}$*: [note different norm!]*

$$\mathrm{Rad}(\mathcal{A} \circ \boldsymbol{x}) \leq \inf_{\varepsilon\in[0,c_{\boldsymbol{x}}/2]} \left\{4\varepsilon + \frac{12}{\sqrt{n}}\int_{\varepsilon}^{c_{\boldsymbol{x}}/2} \sqrt{\log \mathrm{Cov}(\mathcal{A}, \|\cdot\|_{2,\boldsymbol{x}}, \nu)}\, \mathrm{d}\nu\right\}$$

5.1 Applications of the chaining bound for Rademacher complexity

Proposition 5.8 (Rademacher bound for ℓ_∞weights)**.** $\mathcal{A}_\infty := \{x \mapsto w^\top x : \|w\|_\infty \leq 1\}$*, for any* $\boldsymbol{x} := \{x_1, ..., x_n\} \in \mathcal{X}^n$*,*

$$\mathrm{Rad}(\mathcal{A}_\infty \circ \boldsymbol{x}) \leq 12\gamma\frac{\max_i \|x_i\|_1}{\sqrt{n}}\sqrt{d}$$

where $\gamma := \int_0^{1/2} \sqrt{\log(3/\nu)}\, \mathrm{d}\nu$

Proposition 5.9 (Packing bound for binary classifiers). $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathcal{X} \to \{0,1\}\}$, $\mathrm{VCD}(\mathcal{A}) \le \infty$*: for any* $x = \{x_1, ..., x_n\} \in \mathcal{X}^n, p \ge 1$

$$\mathrm{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \le \left(\frac{10}{\varepsilon^p} \log \frac{2\varepsilon}{\varepsilon^p}\right)^{\mathrm{VCD}(\mathcal{A})}$$

Proposition 5.10 (Rademacher bound for binary classifiers). $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathcal{X} \to \{0,1\}\}$, $\mathrm{VCD}(\mathcal{A}) \le \infty$*: for any* $x = \{x_1, ..., x_n\} \in \mathcal{X}^n, p \ge 1$

$$\mathrm{Rad}(\mathcal{A} \circ x) \le 31\sqrt{\mathrm{VCD}(\mathcal{A})/n}$$

6 Sub-gaussians & probability bounds

Proposition 6.1 (Markov's inequality). $X \ge 0$ *is an rv,* $\varepsilon \ge 0$*:*

$$\mathbb{P}(X \ge \varepsilon) \le \frac{\mathbb{E}X}{\varepsilon}$$

Proposition 6.2 (Chernoff bound). X *an* RV, $\lambda \ge 0, \varepsilon \in \mathbb{R}$*:*

$$\mathbb{P}(X \ge \varepsilon) \le \exp(-\lambda\varepsilon)\mathbb{E}\exp(\lambda X)$$

Proposition 6.3 (Optimal Chernoff bound). *given* $\psi^*(x) = \sup_{\lambda \ge 0}(\lambda x - \psi(\lambda))$ *for* $\psi : \mathbb{R}_+ \to \mathbb{R}$*, if* X *is an* RV *st* $\mathbb{E}\exp(\lambda(X - \mathbb{E}X) \le \exp(\psi(\lambda))$ *for all* $\lambda \ge 0$*, then* $\forall\varepsilon \ge 0, \delta \in [0,1]$*:*

$$\mathbb{P}(X - \mathbb{E}X \ge \varepsilon) \le \exp(-\psi^*(\varepsilon))$$

and the same applies for$-X$

Proposition 6.4 (Sums with (Optimised) Chernoff). $X_1, ..., X_n \sim X$ *are iid RVs,* $\mathbb{E}\exp(\lambda(X - \mathbb{E}X) \le \exp(\psi(\lambda))$ *for all* $\lambda \ge 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \ge \varepsilon\right) \le \exp(-n\psi^*(\varepsilon))$$

Definition 6.5 (Sub-Gaussian RV). X is sub-Gaussian if $\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \le \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$ for a variance proxy $\sigma^2 > 0$

Proposition 6.6 (Sub-Gaussian (upper) tail bound). X *sub-Gaussian with* σ^2*.* $\forall\varepsilon \ge 0, \delta \in [0,1]$

$$\mathbb{P}(X - \mathbb{E}X \ge \varepsilon) \le \exp(-\varepsilon^2/(2\sigma^2))$$

Proposition 6.7 (Tail bounds $\implies$ sub-Gaussian). *if* $\forall\varepsilon \ge 0$*,* X *satisfies* $\mathbb{P}(|X - \mathbb{E}X| \ge \varepsilon) \le 2\exp(-\varepsilon^2/(2\sigma^2))$*, then* X *is sub-Gaussian with variance proxy* $\sqrt{8}\sigma^2$ *- i.e.* $\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \le \exp(4\sigma^2\lambda^2)$

Proposition 6.8 (Hoeffding's inequality). *for* $X_1, ..., X_n \sim X$ *iid sub-Gaussian RVs, variance proxy* σ^2*.* $\forall\varepsilon \ge 0, \delta \in [0,1]$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \ge \varepsilon\right) \le \exp(-n\varepsilon^2/(2\sigma^2))$$

Definition 6.9 (Martingale increments).

$$\Delta_i := \mathbb{E}[f(X_1, ..., X_n)|X_1, ..., X_i] - \mathbb{E}[f(X_1, ..., X_n)|X_1, ..., X_{i-1}]$$

$$f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) = \sum_{i-1}^{n} \Delta_i$$

Lemma 6.10 (Azuma's). *if*$\mathbb{E}[\exp(\lambda\Delta_i)|X_1, ..., X_{i-1}] \le \exp(\lambda^2\sigma_i^2/2)$ *for each* $i \in [n]$*, then* $\sum_{i=1}^n \Delta_i$ *is sub-Gaussian with variance proxy* $\sum_{i=1}^n \sigma_i^2$

Definition 6.11 (discrete derivatives). given $f : \mathbb{R}^n \to \mathbb{R}$, for any $x = (x_1, ..., x_n)$:

$$\begin{aligned}\delta_i f(x) &:= \sup_{z\in\mathbb{R}} f(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n) - \inf_{z\in\mathbb{R}} f(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n) \\ \|\delta_i f\|_\infty &:= \sup_{x\in\mathbb{R}^n} |\delta_i f(x)|\end{aligned}$$

Theorem 6.12 (McDiarmid's inequality). *let $X_1, ..., X_n$be independent RVs, then $f(X_1, ..., X_n)$ is sub-Gaussian with variance proxy $\frac{1}{4}\sum_{i=1}^n \|\delta_i f\|_\infty^2$, so*

$$\mathbb{P}(f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) \geq \varepsilon) \leq \exp(-2\varepsilon^2 / \sum_{i=1}^n \|\delta_i f\|_\infty^2)$$

6.1 Applying Hoeffding and McDiarmid to learning

Proposition 6.13. *loss function ℓ bounded in $[0, c]$, $\forall\delta \in [0, 1]$, if $\mathcal{A}$ is* <u>*finite*</u>

$$\mathbb{P}\left(r(A^*) - r(a^*) < c\sqrt{\frac{2\log(2|\mathcal{A}|/\delta}{n}}\right) \geq 1 - \delta$$

Theorem 6.14 (McDiarmid for learning). *if loss function ℓ bounded in $[0, c]$, $\forall\delta \in [0, 1]$*

$$\mathbb{P}\left(r(A^*) - r(a^*) < \mathbb{E}\left[\underbrace{\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\}}_{\textit{Statistics}}\right] + c\sqrt{2\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta$$

$$\mathbb{P}\left(r(A^*) - r(a^*) < 4\mathbb{E}\,\mathrm{Rad}(\mathcal{L} \circ \{Z_1, ..., Z_n\}) + c\sqrt{2\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta$$

Stock-taking

Can bound (in prob), the excess risk of the ERM with Rad, which we can bound in itself, but this is generally with $1/\sqrt{n}$ rates.

7 Bernstein & fast rates

Definition 7.1 (One-sided Bernstein condition). X satisfies the <u>1-sided Bernstein condition</u> with param $b > 0$ if

$$\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp\left(\frac{\lambda^2/2 \cdot \mathrm{Var}(X)}{1 - b\lambda}\right) \quad \text{for any } \lambda \in [0, 1/b)$$

Proposition 7.2 (Bernstein's upper tail bound). *Given X is a 1-sided Bernstein RV with $b > 0$, $\forall\varepsilon \geq 0, \delta \in (0, 1)$, by applying the optimal Chernoff bound,*

$$\mathbb{P}(X - \mathbb{E}X \geq \varepsilon) \leq \exp\left(-\frac{\mathrm{Var}\,X}{b^2}h\left(\frac{b\varepsilon}{\mathrm{Var}\,X}\right)\right) \leq \exp\left(-\frac{\varepsilon^2/2}{\mathrm{Var}\,X + b\varepsilon}\right)$$

$$\mathbb{P}(X - \mathbb{E}X < b\log 1/\delta + \sqrt{2(\mathrm{Var}\,X)\log(1/\delta)}) \geq 1 - \delta$$

where $h : \mathbb{R}_+ \to \mathbb{R}$ is defined by $h(u) = 1 + u - \sqrt{1 + 2u}$, and its inverse is $h^{-1}(u) = u + \sqrt{2u}$

Corollary 7.3 (Bernstein's inequality). *given $X_1, ..., X_n \sim X$ are iid one-sided Bernstein, params $b > 0$ $\forall\varepsilon \geq 0, \delta \in [0, 1]$*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq \varepsilon) \leq \exp\left(-\frac{n \operatorname{Var} X}{b^2} h\left(\frac{b\varepsilon}{\operatorname{Var} X}\right)\right) \leq \exp\left(-\frac{n\varepsilon^2/2}{\operatorname{Var} X + b\varepsilon}\right)$$

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X < \frac{b}{n}\log 1/\delta + \sqrt{2(\operatorname{Var} X)\log(1/\delta)/n}) \geq 1-\delta$$

Which follows from 6.3.

Proposition 7.4 (bounded $\implies$ one-sided Bernstein). *if $X - \mathbb{E}X \leq c$ for a fixed $c > 0$, then X is one-sided Bernstein with $b := c/3$*

Definition 7.5 (2-sided Bernstein condition). with parameter $b > 0$:

$$\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp\left(\frac{\lambda^2/2 \cdot \operatorname{Var}(X)}{1 - b|\lambda|}\right) \quad \text{for any } \lambda \in (-1/b, 1/b)$$

Proposition 7.6 (bounded $\implies$ 2-sided Bernstein). *if $|X - \mathbb{E}X| \leq c$ for a fixed $c > 0$, then X is one-sided Bernstein with $b := c/3$*

Proposition 7.7 (2-sided Bernstein $\implies$ sub-exponential). *a 2-sided Bernstein RV X with $b > 0$ is sub-exponential - i.e. $\exists a, c \geq 0$ st $\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(a\lambda^2)$ for $|\lambda| < 1/c$*

7.1 Fast rates

Consider the binary classification setup: $X_i \in \mathbb{R}^d, Y_i \in \{-1, 1\}$, $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$, loss function is $\ell(a, (x, y)) := \phi(a(x), y)$ for $\phi(\hat{y}, y) := \mathbf{1}_{\hat{y} \neq y}$

So $r(a) = \mathbb{P}(a(X) \neq Y)$, and the Bayes decision rule is $a^{**}(x) = \arg\max_{\hat{y} \in \mathcal{Y}} \mathbb{P}(Y = \hat{y}|X = x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{if } \eta(x) \leq 1/2 \end{cases}$, where the regression function is $\eta(x) = \mathbb{P}(Y = 1|X = x)$

Theorem 7.8. *For any classifier $a \in \mathcal{B}$:*

$$r(a) - r(a^{**}) = \mathbb{E}\left[|2\eta(X) - 1| \cdot \mathbf{1}_{a(X) \neq a^{**}(X)}\right]$$
$$r(a^{**}) = \mathbb{E}\min\{\eta(X), 1 - \eta(X)\} \leq 1/2$$

By conditional expectation...

Definition 7.9 (Massart's noise condition). A problem satisfies this condition if $\exists \gamma \in (0, 1/2]$ st $\forall x \in \mathbb{R}^d$

$$\left|\eta(x) - \frac{1}{2}\right| \geq \gamma$$

The noise-less case is $\gamma = 1/2$, so $\eta = 0$ or 1, and the information-free case would be $\gamma = 0$

Theorem 7.10 (Fast rate for binary classification w/ Massart's noise). *Let $a^{**} \in \mathcal{A}$,so $a^* = a^{**}$, and assume Massart's noise condition holds. Then*

$$\mathbb{P}\left(r(A^*) - r(a^*) \leq \frac{\log(|\mathcal{A}|/\delta)}{\gamma n}\right) \geq 1 - \delta$$

Definition 7.11 (Tsybakov's noise condition). $\exists \alpha \in (0, 1), \beta > 0, \gamma \in (0, 1/2]$ st $\forall t \in [0, \gamma]$

$$\mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq \beta t^{\frac{\alpha}{1-\alpha}}$$

As $\alpha \to 0$ $t^{\alpha/(1-\alpha)} \to 1$, so the condition is void. $\alpha \to 1$ recovers Massart's condition, so $\alpha \in (0, 1)$ will "interpolate" between $1/\sqrt{n}$ and $1/n$.

Proposition 7.12. *Given Tsybakov's noise condition holds, there is a constant c (depending on α, β, γ) st $\forall a \in \mathcal{A}$*

$$\mathbb{P}[a(X) \neq a^{**}(X)] \leq c(r(a) - r(a^{**}))^\alpha$$

Theorem 7.13 (Tsybakov: interpolation). *Given $a^{**} \in \mathcal{A}$, so $a^* = a^{**}$, and the conditions of 7.12*

$$\mathbb{P}\left(r(A^*) - r(a^*) \leq c\left(\frac{\log(|\mathcal{A}|/\delta}{n}\right)^{\frac{1}{2-\alpha}}\right) \geq 1 - \delta$$

8 Convexity

Working in the same binary classification setup as in 7.1, with the true loss, defined as $\phi(\hat{y}, y) := \varphi^*(\hat{y}y)$, where $\varphi^*(u) := \mathbf{1}_{u \leq 0}$

Convex function, set definitions as standard

Definition 8.1 (Convex loss surrogate)**.** $\varphi : \mathbb{R} \to \mathbb{R}_+$ which is convex, non-increasing and $\varphi(0) = 1$. Thus, it is an upper bound on the true loss, as $\varphi^*(u) \leq \varphi(u)$. Some examples include:

- **exponential loss** $\varphi(u) := \exp(-u)$
- **hinge loss** $\varphi(u) := \max\{1 - u, 0\}$
- **logistic loss** $\varphi(u) := \log_2(1 + \exp(-u))$

$\mathcal{B}_{\textsf{soft}} := \{a : \mathbb{R}^d \to \mathbb{R}\}$ instead of to $\{-1, 1\}$. The minimisation problems are the same, and they are now convex. The corresponding hard classifier is $\operatorname{sign}(a)$. Common choices:

- **linear functions w/ convex parameter space:** $\mathcal{A}_{\textsf{soft}} := \{a(x) = w^\top x + b : w \in S_1, b \in S_2\}$, where $S_1 \subseteq \mathbb{R}^d, S_2 \subseteq \mathbb{R}$ are both convex
- **Boosting/majority votes** $\mathcal{A}_{\textsf{soft}} := \{a(x) = \sum_{i=1}^m w_i h_i(x) : (w_1, ..., w_m) \in \Delta_m\}$, where Δ_m is the simplex, and $h_1, ..., h_m$ are the base classifiers $\mathbb{R}^d \to \mathbb{R}$

$$r_\varphi(a) - r_\varphi(a^{**})$$

where $r_\varphi(a) := \mathbb{E}[\varphi(a(X)Y)]$ (and $R_\varphi(a) := \frac{1}{n}\sum_{i=1}^n \varphi(a(X_i) \cdot Y_i)$. This φ-risk is not necessarily related to the normal excess risk for the hard classifier $\operatorname{sign}(a)$.

Lemma 8.2 (Zhang's)**.** *let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a convex loss surrogate, and for any $\tilde{\eta} \in [0, 1]$, $\tilde{a} \in \mathbb{R}$ define*

$$H_{\tilde{\eta}}(\tilde{a}) := \varphi(\tilde{a})\tilde{\eta} + \varphi(-\tilde{a})(1 - \tilde{\eta}), \qquad \tau(\tilde{\eta}) := \inf_{\tilde{a} \in \mathbb{R}} H_{\tilde{\eta}}(\tilde{a}).$$

If there exist $c > 0, \nu \in [0, 1]$ st

$$\left|\tilde{\eta} - \frac{1}{2}\right| \leq c(1 - \tau(\tilde{\eta}))^\nu \text{ for any } \tilde{\eta} \in [0, 1]$$

Then $\forall a : \mathbb{R}^d \to \mathbb{R}$ we have

$$\underbrace{r(\operatorname{sign}(a)) - r(a^{**})}_{\text{excess risk hard classifier}} \leq 2c \underbrace{\left(r_\varphi(a) - r_\varphi(a^{**}_\varphi)\right)^\nu}_{\text{excess }\varphi\text{-risk soft classifier}}$$

Definition 8.3 (Convexity definitions)**.** The convex hull is as expected, the epigraph is

$$\operatorname{epi}(f) := \{(x, t) \in \mathcal{D} \times \mathbb{R} : f(x) \leq t\}$$

Proposition 8.4 (Equivalent optimisations)**.** *(also for* $\max$ *instead of* $\min$*):*

$$\min_{t \in \mathcal{T}} c^\top t = \min_{t \in \operatorname{conv}(\mathcal{T})} c^\top t$$

Proposition 8.5 (Epigraph in optimisation)**.**

$$\min_{x \in \mathcal{D}} f(x) = \min_{(x,t) \in \operatorname{epi}(f)} t$$

Remark 8.6. Thus, combining these two propositions, minimising any function is equivalent to minimising over the convex hull of the epigraph of the convex function $(x, t) \to t$

Definition 8.7 (Subgradient)**.** $g \in \mathbb{R}^d$is a subgradient of $f : \mathcal{C} \subseteq \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathcal{C}$ if

$$f(x) - f(y) \leq g^\top (x - y) \text{ for any } y \in \mathcal{C}.$$

$\partial f(x)$ is the set of subgradients of f at x

Theorem 8.8 (convexity and subgradients). *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be convex. If $\forall x \in \mathcal{C}$, $\partial f(x) \neq \emptyset$, then f is convex. Conversely, if f is convex then for all $x \in \text{int}(\mathcal{C})$, $\partial f(x) \neq \emptyset$. The derivative of f is always a subgradient, if it exists.*

Proposition 8.9 (1st order optimality condition). *If f is convex, and $\mathcal{C}$ is a closed set and f is differentiable on it, then*

$$x^* \in \arg\min_{x \in \mathcal{C}} f(x) \iff \nabla f(x^*)^\top (x^* - y) \leq 0 \text{ for any } y \in \mathcal{C}$$

Definition 8.10 (Convexity, smoothness and Lipschitz).

- **convex**: $f(x) - f(y) \leq \nabla f(x)^\top (x-y)$ for any $x, y \in \mathbb{R}^d$
- α-**strongly convex**: $\exists \alpha > 0$ st $f(x) - f(y) \leq \nabla f(x)^\top (x-y) - \frac{\alpha}{2}\|x-y\|_2^2$ for any $x, y \in \mathbb{R}^d$
- β-**smooth**: $\exists \beta > 0$ st $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x-y\|_2$ for any $x, y \in \mathbb{R}^d$
- γ-**Lipschitz**: $\exists \gamma \geq 0$ st $|f(x) - f(y)| \leq \gamma |x-y|$ for any $x, y \in \mathbb{R}^d$, or $\exists \gamma > 0$ st $\|\nabla f(x)\|_2 \leq \gamma$ for a ny $x \in \mathbb{R}^d$

Definition 8.11 (8.10, when twice differentiable). Note $M \preceq M_2$ if the largest eigenvalue of M is less than the smallest eigenvalue of M_2

- **convex**: $\nabla^2 f(x) \succeq 0$ for any $x \in \mathbb{R}^d$
- α-**strongly convex**: $\exists \alpha > 0$ st $\nabla^2 f(x) \succeq \alpha I$ for any $x \in \mathbb{R}^d$
- β-**smooth**: $\exists \beta > 0$ st $\nabla^2 f(x) \preceq \beta I$ for any $x \in \mathbb{R}^d$
- γ-**Lipschitz**: same as above

Definition 8.12 (Properties of common convex loss surrogates).

loss surrogate	strongly convex?	smooth?	Lipschitz?
exponential	N	N	N
hinge	N	N	Y - $\gamma = 1$
logistic	N	Y	Y
least square	Y	Y	N

(note sure what least squares is here - maybe the more general $(a(x) - y)^2$?

TODO: problem 3.5

9 Oracle model & gradient descent

Definition 9.1 (Oracle model). We have an oracle for some property of the function (e.g. subgradient), and measure the complexity of an algorithm using the oracle by how many times it calls the oracle.

Definition 9.2 (Projection operator). $\forall y \in \mathbb{R}^d$

$$\Pi_{\mathcal{C}}(y) := \arg\min_{x \in \mathcal{C}} \|x - y\|_2$$

Proposition 9.3 (Non-expansivity). *given $x \in \mathcal{C}, y \in \mathbb{R}^d$*

$$(\Pi_{\mathcal{C}}(y) - x)^\top (\Pi_{\mathcal{C}}(y) - y) \leq 0$$

which implies

$$\|\Pi_{\mathcal{C}}(y) - x\|_2 \leq \|y - x\|_2$$

Algorithm 1 Projected Subgradient Method

```
Input: $x_1, \{\eta_s\}_{s \geq 1}$, stopping time $t$
for $j = 1, ..., t$ do
    $\tilde{x}_{s+1} \leftarrow x_s - \eta_s g_s$, where $g_s \in \partial f(x_s)$
    $x_{s+1} \leftarrow \Pi_{\mathcal{C}}(\tilde{x}_{s+1})$
end for
```

Theorem 9.4 (PSubGD: Lipschitz). *Given f is convex, and $\forall x \in \mathcal{C}\ \forall g \in \partial f(x)\ \|g\|_2 \leq \gamma$, and assume $\|x_1 - x^*\| \leq b$. THen the projected subgradient method, with $\eta_s = \eta := b/(\gamma\sqrt{t})$ satisfies*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{\gamma b}{\sqrt{t}} \text{ and } f(x^{'}) - f(x^*) \leq \frac{\gamma b}{\sqrt{t}},$$

where $x' \in \arg\min_{x \in \{x_1,\ldots,x_t\}} f(x)$.

If $\tilde{b}$ is an upper bound on the diameter of $\mathcal{C}$, so $\sup_{x,y\in\mathcal{C}} \|x-y\|_2 \leq \tilde{b}$, given $\tilde{\eta}_s := \frac{\tilde{b}}{\gamma\sqrt{t}}$, then

$$f\left(\frac{1}{\sum_{s=\lceil t/2\rceil+1}^{t}\eta_s}\sum_{s=\lceil t/2\rceil+1}^{t}\eta_s x_s\right) - f(x^*) \leq 2(1+\log 2) \times \frac{\gamma\tilde{b}}{\sqrt{t}}$$

(plus the same result about the argmin x' again.

This is therefore slow rate convergence: $1/\sqrt{t}$.

Theorem 9.5 (PGD: Smooth). *Let f be convex and β-smooth on $\mathcal{C}$. With $\eta = 1/\beta$, PGD satisfies:*

$$f(x_t) - f(x^*) \leq \frac{3\beta\|x_1 - x^*\|_2^2 + f(x_1) - f(x^*)}{t}$$

Which is a fast rate!

Note that this is gradient descent not subgradient descent, because smoothness requires differentiability.

[todo for proof: eqs. 9.6 and 9.7

Theorem 9.6 (GD: Smooth and Strongly convex). *Let f be α-strongly convex and β-smooth on $\mathbb{R}^d$. Then gradient descent with $\eta = 1/\beta$ satisfies*

$$f(x_t) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^{t-1}(f(x_1) - f(x^*))$$

Remark 9.7 (Comparison of (P)(Sub)GD results). Given $b \geq \|x_1 - x^*\|_2$ (and is at worst, the diameter, if finite), and $c \geq f(x_1) - f(x^*)$

	γ-Lipschitz	β-smooth
convex	$\frac{\gamma b}{\sqrt{t}}$ (9.4)	$\frac{\beta b^2 + c}{t}$ 9.5
α-strongly convex	$\frac{\gamma^2}{\alpha t}$(see ref 2. in ch 9)	$\exp(-t\alpha/\beta)c$ 9.6

Gives the convergence rates for each case (up to $O(...)$ terms). If we want the oracle complexity to achieve a given accuracy ε, rearrange these for t.

Remark 9.8 (Lower bounds on convergence rates). The following table (from ref [2]), gives lower bounds of $f(x') - f(x_0)$, st there exists an $f$achieving them.

	γ-Lipschitz	β-smooth
convex	$\Omega(\gamma a/(1+\sqrt{t})$	$\Omega(b^2\beta/(t+1)^2)$
α-strongly convex	$\Omega(\gamma^2/(at)$	$\Omega(a\tilde{b}\exp(-t\sqrt{\alpha/\beta}))$

Application to classification with ℓ_2 constraints

Linear predictors with convex constraints, so $\mathcal{A} = \{x \mapsto a(x) = w^\top x : w \in \mathcal{W} \subseteq \mathbb{R}^d\}$, where $\mathcal{W}$ is a convex set contained in a ball of radius $c^{\mathcal{W}}$.

Features $X \in \mathbb{R}$, binary labels $Y \in \{-1, 1\}$

Using the same decomposition as in 1, using w^*, W^*, W as functions: for any $W \in \mathcal{W}$

$$\leq r(W) - r(w^*) \leq \underbrace{(R(W) - R(W^*))}_{\text{Optimisation}} + \underbrace{\sup_{w\in\mathcal{W}}\{r(w) - R(w)\} + \sup_{w\in\mathcal{W}}\{r(w) - R(w)\}}_{\text{Statistics}}$$

Under the assumptions that φ is γ-Lipschitz and $\mathcal{X}$ is contained in a ball of radius $c^{\mathcal{X}}$

$\mathbb{E}\sup_{w\in\mathcal{W}}\{r(w)-R(w)\} \leq 2\mathbb{E}\,\mathrm{Rad}(\mathcal{L}\circ\{Z_1,...\} \leq 2\gamma\mathbb{E}\,\mathrm{Rad}(\mathcal{A}\circ\{Z_1,...,\}) \leq \frac{2\gamma c^{\mathcal{X}}c^{\mathcal{W}}}{\sqrt{n}}$

So we can bound the Statistics term by $\frac{4\gamma c^{\mathcal{X}}c^{\mathcal{W}}}{\sqrt{n}}$, and the Optimisation term, we can sort as R is $(c^{\mathcal{X}}\gamma)$-Lipschitz

$$|R(w)-R(u)| \leq \frac{1}{n}\sum_{i=1}^{n}|\varphi(w^\top X_i \times Y_i) - \varphi(u^\top X_i \times Y_i) \leq \frac{\gamma}{n}\sum_{i=1}^{n}|Y_i(w-u)^\top X_i| \leq \gamma c^{\mathcal{X}}\|w-u\|_2$$

By C-S, $|Y_i| = 1$

So we use projective subgradient descent 9.4 to get the Optimisation term bounded by $\dfrac{2c^{\mathcal{X}}c^{\mathcal{W}}\gamma}{\sqrt{t}}$ (using the diameter of the weights being finite).

10 Non-Euclidean Gradient Descent & Mirror descent

We have issues if the Lipschitz constant of the loss varies with d or the Statistics bound does - e.g. when using the simplex/ℓ_∞setup -e.g. in 3.4, where the empirical risk is also $\sqrt{d}\times\cdots$ Lipschitz. This occurs in many/all? non-Euclidean setups (i.e. not the 2 norm).

The definitions of strongly convex/smooth/Lipschitz can all be restated in terms of arbitrary norms.

In a Hilbert space [i.e. inner products!], the methods in 9 work, as even though the derivatives are more correctly Fréchet derivatives (below) and thus in the dual space, the Riesz representation theorem still allows us to consider these as vectors.

We will use Hilbert spaces in $\mathbb{R}^d$, with various norms, but also allowing the 2-norm inner product, but without any special meaning.

Definition 10.1 (Fréchet derivative). Given normed spaces $(\mathcal{D},\|\cdot\|_{\mathcal{D}}),(\mathcal{D}',\|\cdot\|_{\mathcal{D}'})$ $T:\mathcal{C}\subseteq\mathcal{D}\to\mathcal{D}'$ (for open $\mathcal{C}$) is Fréchet differentiable at $d\in\mathcal{C}$ if $\exists$a bounded linear operator $D_dT:\mathcal{D}\to\mathcal{D}'$, the Fréchet derivative at d, st

$$\lim_{h\to 0}\frac{\|T(d+h)-T(d)-D_dT(h)\|_{\mathcal{D}'}}{\|h\|_{\mathcal{D}}} = 0$$

[n.b. h is not a real number, but an element of $\mathcal{D}$]

Algorithm 2 Projected Mirror Descent

```
Input: $x_1, \{\eta_s\}_{s\geq 1}$, stopping time $t$
for $j = 1,...,t$ do
    map $x_s$ to $\nabla\Phi(x_s)$
    $\nabla\Phi(\tilde{x}_{s+1}) \leftarrow \nabla\Phi(x_s) - \eta_s g_s$, where $g_s \in \partial f(x_s)$
    map $\nabla\Phi(\tilde{x}_{s+1})$ back to $\tilde{x}_{s+1}$
    $x_{s+1} \leftarrow \Pi_{\mathcal{C}}^{\Phi}(\tilde{x}_{s+1})$
end for
```

Definition 10.2 (Mirror map). $\mathcal{D}\subseteq\mathbb{R}^d$ convex open set, $\mathcal{C}\subseteq\overline{D}, \mathcal{C}\cap\mathcal{D}\neq\emptyset$. $\Phi:\mathcal{D}\to\mathbb{R}$ is a **mirror map** if:

1. Φ is strictly convex and differentiable
2. $\nabla\Phi:\mathcal{D}\to\mathbb{R}^d$ is a surjective map
3. the gradient diverges on the boundary of $\mathcal{D}$- i.e. $\lim_{x\to\delta D}\|\nabla\Phi(x)\| = \infty$

idea: "$\nabla\Phi$ maps the space into its dual, where we can add gradients (as they are functionals), and then allows us to map backwards"

Definition 10.3 (Bregman divergence). For a differentiable function $g:\mathbb{R}^d\to\mathbb{R}$:

$$D^g(x,y) := g(x) - g(y) - \nabla g(y)^\top(x-y)$$

Definition 10.4 (Bregman projection)**.** given Φ is a mirror map:

$$\Pi_{\mathcal{C}}^{\Phi}(y) = \underset{x \in \mathcal{C} \cap \mathcal{D}}{\arg\min} \, D^{\Phi}(x, y)$$

N.B. properties 1 and 3 of the mirror map ensure this projection exists and is unique.

Example 10.5 (PSubGD from Euclidean balls)**.** We get projected subgradient descent again if we take $\mathcal{D} = \mathbb{R}^d$ and $\Phi(x) = \frac{1}{2}\|x\|_2^2$, so the Bregman divergence is

$$D^{\Phi}(x, y) = \cdots = \frac{1}{2}\|x - y\|_2^2$$

So the Bregman projection is the same as the Euclidean projection.

Example 10.6 (Exponential GD from negative entropy)**.** if $\mathcal{C} = \Delta_d$ is the simplex, and the mirror map is the negative entropy

$$\Phi(x) = \sum_{i=1}^{d} x_i \log x_i,$$

and $\mathcal{D} = \{x \in \mathbb{R}^d : x_i > 0 \text{ for all } i \in [d]\}$, then $\nabla\Phi(x) = 1 + \log x$ (componentwise $\log$), the mirror update is

$$\log(\tilde{x}_{s+1}) = \log(x_s) - \eta g_s,$$

which is equivalent to

$$\tilde{x}_{s+1} = x_s \exp(-\eta g_s)$$

And the Bregman divergence is just the K-L divergence. Using Pinsker's inequality, Φ is 1-strongly convex wrt the $\|\cdot\|_1$ norm, and using KKT conditions the Bregman projection is just a normalisation step: $\Pi_{\mathcal{C}}^{\Phi}(y) = \frac{y_i}{\sum_i y_i}$

Proposition 10.7 (Pinsker's inequality)**.** *if $x, y \in \Delta_d$, then*

$$\|x - y\|_{tv} := \frac{1}{2}\sum_{i=1}^{d} |x_i - y_i| \leq \sqrt{\frac{1}{2}\sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}} = \sqrt{\frac{1}{2} D_{K\text{-}L}(x \parallel y)}$$

Proposition 10.8 (Bregman divergence and derivative identity)**.** *for any differentiable $g : \mathbb{R}^d \to \mathbb{R}$:*

$$(\nabla g(x) - \nabla g(y))^{\top} (x - z) = D^g(x, y) + D^g(z, x) - D^g(z, y)$$

Proposition 10.9 (Non-expansivity)**.** *let $x \in \mathcal{C} \cap \mathcal{D}, y \in \mathcal{D}$, then*

$$\left(\nabla\Phi(\Pi_{\mathcal{C}}^{\Phi}(y) - \nabla\Phi(y)\right)^{\top} (\Pi_{\mathcal{C}}^{\Phi}(y) - x) \leq 0$$

which implies $D^{\Phi}(x, \Pi_{\mathcal{C}}^{\Phi}(y)) + D^{\Phi}(\Pi_{\mathcal{C}}^{\Phi}(y), y) \leq D^{\Phi}(x, y)$ and

$$D^{\Phi}(x, \Pi_{\mathcal{C}}^{\Phi}(y)) \leq D^{\Phi}(x, y)$$

Theorem 10.10 (PMD: Lipschitz)**.** *Given:*

- *f convex, γ-Lipschitz wrt $\|\cdot\|$*
- *Φ an α-strongly convex mirror map on $\mathcal{C} \cap \mathcal{D}$ wrt $\|\cdot\|$*
- *assume $x_1 \in \arg\min_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x)$*

The projected mirror descent algorithm with $\eta_s := \eta := \frac{c}{\gamma}\sqrt{\frac{2\alpha}{t}}$ satisfies

$$f(\frac{1}{t}\sum_{s=1}^{t} x_s) - f(x^*) \leq c\gamma\sqrt{\frac{2}{\alpha t}}$$

, where $c^2 = \sup_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - \Phi(x_1)$

Proof. TIDY up!!!!

- $f(x_s) - f(x^*) \le g_s^T(x_s - x^*) = (\frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1})))^\top(x_s - x^*)$
    - $\le$ by def of subgradient and convex,
    - $=$ by $g_s = \frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))$, by the update step
- $= \frac{1}{\eta}\left(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x^*, x_s) - D^\Phi(x^*, \tilde{x}_{s+1})\right)$ by prop 10.9
- $\le \frac{1}{\eta}\left(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x^*, x_s) - D^\Phi(x^*, x_{s+1})\right)$ by non-expansivity of D^Φ(prop 10.10)

we have:

- $D^\Phi(x_s, \tilde{x}_{s+1}) = \Phi(x_s) - \Phi(\tilde{x}_{s+1}) - \nabla\Phi(x_{s+1})^T(x_s - \tilde{x}_{s+1})$ by definition of D^Φ
- $\le (\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))^\top (x_s - \tilde{x}_{s+1}) - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2$
    - since $\Phi(\tilde{x}_{s+1}) \ge \Phi(x_s) + \nabla\Phi(x_s)^T(\tilde{x}_{s+1} - x_s) + \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2$ by α-strong convexity of Φ
- $= \eta g_s^\top(x_s - \tilde{x}_{s+1}) - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2$ by def of g_s
- $\le \eta\gamma\|x_s - \tilde{x}_{s+1}\| - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2$ by Holder's inequality, and then $\|g_s\|_* \le \gamma$ by Lipschitz
- $\le \frac{\eta^2\gamma^2}{2\alpha}$ by the inequality $az - bz^2 \le \max_{z\in\mathbb{R}}(az - bz^2) = a^2/4b$

Putting it all together:

- $f(\frac{1}{t}\sum_{s=1}^t x_s) - f(x^*) \le \frac{1}{t}\sum_{s=1}^t(f(x_s) - f(x^*))$ by convexity
- $\le \frac{1}{\eta t}\sum_{s=1}^t \left(D^\Phi(x^*, x_s) - D^\Phi(x^*, x_{s+1})\right) + \frac{\eta^2\gamma^2}{2\alpha}$ by both arguments above
- $= \frac{1}{\eta t}\left(D^\Phi(x^*, x_1) - D^\Phi(x^*, x_{t+1})\right) + \frac{\eta^2\gamma^2}{2\alpha}$ by telescoping
- $\le \frac{1}{\eta t}D^\Phi(x^*, x_1) + \frac{\eta^2\gamma^2}{2\alpha}$ since $D^\Phi(x^*, x_{t+1})$ is non-negative as Φ convex
- $\le \frac{1}{\eta t}\sqrt{c} + \frac{\eta^2\gamma^2}{2\alpha}$ as $D^\Phi(x^*, x_1) = \Phi(x^*) - \Phi(x_1) - \nabla\Phi(x_1)^\top(x^* - x_1) \le \Phi(x^*) - \Phi(x_1) \le \sqrt{c}$
    - note that the optimality condition in 8.10 gives us that $\nabla\Phi(x_1)^\top(x^* - x_1) \ge 0$ by the assumption that $x_1 \in \arg\min_{x\in\mathcal{C}\cap\mathcal{D}}\Phi(x)$

Then plug in η as required. □

10.1 Application to learning

Same setup as 9, except that

$$\mathcal{A}_\Delta := \{x \mapsto a(x) = w^\top x : w \in \Delta_d\}$$

So the risk minimisation/ERM are all standard.

Choose, as mirror map, the negative entropy $\Phi(w) = \sum_{i=1}^d w_i \log w_i$

so the starting point (as we require a minimum) is

$$w_1 = \frac{1}{d}\mathbf{1}$$

and the constant c^2 is $\log d$.

Given 10.6, $\alpha = 1$ for Φ

and by Holder's inequality (note using two different norms here), $|R(w) - R(u)| \le \cdots \gamma_\varphi c_\infty^{\mathcal{X}}\|w - u\|_1$, so the empirical risk is $\gamma_\varphi c_\infty^{\mathcal{X}}$-Lipschitz wrt the $\|\cdot\|_1$ norm.

Thus, applying the projected mirror descent with step size η for t steps

$$\eta = \frac{c}{\gamma_\varphi c_\infty^{\mathcal{X}}}\sqrt{\frac{2\alpha}{t}} = \frac{1}{\gamma_\varphi c_\infty^{\mathcal{X}}}\sqrt{\frac{2\log d}{t}}$$

(since $c^2 = \log d, \alpha = 1$), we get the

$$\text{Optimisation}_\Delta := R(\overline{W}_t) - R(W_\Delta^*) \leq c^{\mathcal{X}} c^{\mathcal{W}} \sqrt{\frac{2\log d}{t}}$$

which perfectly matches the bound on $\mathbb{E}\text{Statistics}_\Delta$.

11 Stochastic Oracle model

Definition 11.1 (First order stochastic oracle)**.** given $x \in \mathcal{C}$, the oracle returns an RV G st $\mathbb{E}G \in \partial f(x)$, and if $X \in \mathcal{C}$ is an RV, it returns G st $\mathbb{E}[G|X] \in \partial f(X)$

Algorithm 3 Projected Stochastic Subgradient Descent

```
Input: $X_1, \{\eta_s\}_{s\geq 1}$, stopping time $t$
for $j = 1, ..., t$ do
    $\tilde{X}_{s+1} \leftarrow X_s - \eta_s G_s$, where $\mathbb{E}[G_s|X_s] \in \partial f(X_s)$
    $X_{s+1} \leftarrow \Pi_{\mathcal{C}}(\tilde{X}_{s+1})$
end for
```

Theorem 11.2 (Proj. SGD)**.** *Given f is convex, $\mathbb{E}[\|G_s\|_2^2] \leq \gamma^2$ for any $s \in [t]$ and $\mathbb{E}\|X_1 - x^*\|_2^2 \leq b^2$, Then the projected subgradient algorithm with $\eta_s := \eta := \frac{b}{\gamma\sqrt{t}}$ satisfies*

$$\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t} X_s\right) - f(x^*) \leq \frac{\gamma b}{\sqrt{t}}$$

Proof. [tidy up!]

- $f(X_s) - f(x^*) \leq \delta f(X_s)^T(X_s - x^*) = \mathbb{E}[G_s|X_s]^T(X_s - x^*) = \mathbb{E}[G_s^T(X_s - x^*)|X_s]$
    - $\leq$ by def of subgradient and convex,
    - $=$ by cond exp
- by theorem 9.3, $G_s^T(X_s - x^*) \leq \frac{1}{2\eta}\left(\|X_s - x^*\|_2^2 - \|X_{s+1} - x^*\|_2^2\right) + \frac{\eta}{2}\|G_s\|_2^2$
- $\mathbb{E}f(X_s) - f(x^*) = \mathbb{E}[\mathbb{E}[G_s^T(X_s - x^*)|X_s]] = \mathbb{E}[G_s^T(X_s - x^*)] \leq \frac{1}{2\eta}\left(\mathbb{E}\|X_s - x^*\|_2^2 - \mathbb{E}\|X_{s+1} - x^*\|_2^2\right) + \frac{\eta}{2}\mathbb{E}\|G_s\|_2^2$
    - by tower proerty, using points above
- $\frac{1}{t}\sum_{s=1}^{t}(\mathbb{E}f(X_2) - f(x^*)) \leq \frac{1}{2\eta t}(\mathbb{E}\|X_1 - x^*\|_2^2 - \mathbb{E}\|X_{t+1} - x^*\|_2^2) + \frac{\eta}{2}\gamma^2$
    - by telescoping sums & asusmption $\mathbb{E}[\|G_s\|_2^2] \leq \gamma^2$,
- above $\leq \frac{1}{2\eta t}(b^2 + 0) + \frac{\eta}{2}\gamma^2$
    - by assumption $\mathbb{E}\|X_1 - x^*\|_2^2 \leq b^2$, $-\mathbb{E}\|X_{t+1} - x^*\|_2^2$ term non-negative
- then plug in η definition
- use Jensen's for $f(\frac{1}{t}\sum_{s=1}^{t} X_s) \leq \frac{1}{t}\sum_{s=1}^{t} f(X_s)$

□

Algorithm 4 Projected Stochastic Mirror Descent

```
Input: $X_1, \{\eta_s\}_{s\geq 1}$, stopping time $t$
for $j = 1, ..., t$ do
    map $X_s$ to $\nabla\Phi(X_s)$
    $\nabla\Phi(\tilde{X}_{s+1}) \leftarrow \nabla\Phi(X_s) - \eta_s G_s$, where $\mathbb{E}[G_s|X_s] \in \partial f(X_s)$
    map $\nabla\Phi(\tilde{X}_{s+1})$ back to $\tilde{X}_{s+1}$
    $X_{s+1} \leftarrow \Pi^{\Phi}_{\mathcal{C}}(\tilde{X}_{s+1})$
end for
```

Theorem 11.3 (Projected stochastic mirror descent)**.** *Given*

- *f is convex,*
- *$\mathbb{E}[\|G_s\|_2^2] \leq \gamma^2$ for any $s \in [t]$,*
- *Φ is an α-strongly convex mirror map on $\mathcal{C} \cap \mathcal{D}$ wrt $\|\cdot\|$,*
- *and $X_1 := x_1 \in \arg\min_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$*

Then projected mirror descent with $\eta_s := \eta := \frac{c}{\gamma}\sqrt{\frac{2\alpha}{t}}$ satisfies

$$\mathbb{E}f\left(\frac{1}{t}\sum_{s=1}^{t} X_s\right) - f(x^*) \leq c\gamma\sqrt{\frac{2}{\alpha t}},$$

where $c^2 = \sup_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x) - \Phi(x_1)$

Proof. TODO: see PS 3.4

- $f(X_s) - f(x^*) \leq \delta f(X_s)^\top (X_s - x^*) = \mathbb{E}[G_s|X_s]^\top (X_s - x^*)$
  - by definition of the subgradient and that G_s is an unbiased estimator of a subgradient
  - and further by the update step
- $= \mathbb{E}[\frac{1}{\eta}(\nabla\Phi(X_s) - \nabla\Phi(\tilde{X}_{s+1})))(X_s - x^*)|X_s]$
  - by the definition of the update step, and because $X_s = \mathbb{E}[X_s|X_s]$
  - note that $\tilde{X}_{s+1}$ is not simply a function of X_s, so we cannot take it ou
- $\leq \mathbb{E}\frac{1}{\eta}\left(D^\Phi(X_s, \tilde{X}_{s+1}) + D^\Phi(x^*, X_s) - D^\Phi(x^*, X_{s+1})\right)|X_s]$
  - by proceeding as in the proof of theorem 11.1, which is unaffected by
- MISSING
- then plug in η definition
- final step: use Jensen's for $f(\frac{1}{t}\sum_{s=1}^{t} X_s) \leq \frac{1}{t}\sum_{s=1}^{t} f(X_s)$

rest:

- $f(X_s) - f(x^*) \leq \delta f_s^T (X_s - x^*) = (\frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1})))^\top (x_s - x^*)$
  - $\leq$ by def of subgradient and convex,
  - $=$ by $g_s = \frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))$, by the update step
- $= \frac{1}{\eta}\left(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x^*, x_s) - D^\Phi(x^*, \tilde{x}_{s+1})\right)$ by prop 10.9
- $\leq \frac{1}{\eta}\left(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x^*, x_s) - D^\Phi(x^*, x_{s+1})\right)$ by non-expansivity of D^Φ(prop 10.10)
- $f(X_s) - f(x^*) \leq \delta f(X_s)^T (X_s - x^*) = \mathbb{E}[G_s|X_s]^T (X_s - x^*) = \mathbb{E}[G_s^T (X_s - x^*)|X_s]$
  - $\leq$ by def of subgradient and convex,

  - $=$ by cond exp
- by theorem 9.3, $G_s^T(X_s - x^*) \le \frac{1}{2\eta}\left(\|X_s - x^*\|_2^2 - \|X_{s+1} - x^*\|_2^2\right) + \frac{\eta}{2}\|G_s\|_2^2$
- $\mathbb{E}f(X_s) - f(x^*) = \mathbb{E}[\mathbb{E}[G_s^T(X_s - x^*)|X_s]] = \mathbb{E}[G_s^T(X_s - x^*)] \le \frac{1}{2\eta}\left(\mathbb{E}\|X_s - x^*\|_2^2 - \mathbb{E}\|X_{s+1} - x^*\|_2^2\right) + \frac{\eta}{2}\mathbb{E}\|G_s\|_2^2$
  - by tower proerty, using points above
- $\frac{1}{t}\sum_{s=1}^{t}(\mathbb{E}f(X_2) - f(x^*)) \le \frac{1}{2\eta t}(\mathbb{E}\|X_1 - x^*\|_2^2 - \mathbb{E}\|X_{t+1} - x^*\|_2^2) + \frac{\eta}{2}\gamma^2$
  - by telescoping sums & asusmption $\mathbb{E}[\|G_s\|_2^2] \le \gamma^2$,
- above $\le \frac{1}{2\eta t}(b^2 + 0) + \frac{\eta}{2}\gamma^2$
  - by assumption $\mathbb{E}\|X_1 - x^*\|_2^2 \le b^2$, $-\mathbb{E}\|X_{t+1} - x^*\|_2^2$ term non-negative

□

11.1 Application to Learning

Remark 11.4 (Passes through data). We can do SGD with 1 pass through the data, using each point once to calculate a subgradient estimator. We can't deterministically reuse points, as then the subgradient estimators won't be unbiased. This works well for online learning. Theorem 11.1 applied to this gives us

$$\mathbb{E}r(\overline{W}_t) - r(w_2^*) \le \frac{2c_2^{\mathcal{X}}c_2^{\mathcal{W}}\gamma_\varphi}{\sqrt{t}}$$

Or we can do multiple passes by using random IID indices I_j that take values in $\{1, ..., n\}$, as the subgradient estimators are still unbiased, but do not estimate the risk, instead the empirical risk:

$$\mathbb{E}[\partial_w\varphi(W_s^\top X_{I_{s+1}}Y_{I_{s+1}})|S, W_s) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\partial_w\varphi(W_s^\top X_iY_i)|S, W_s) = \partial R(W_s).$$

Thus, we get for any t:

$$\mathbb{E}\texttt{Optimisation}_2 = \mathbb{E}R(\overline{W}_t) - R(W_2^*) \le \frac{2c_2^{\mathcal{X}}c_2^{\mathcal{W}}\gamma_\varphi}{\sqrt{t}}$$

Definition 11.5 (Explicit regularisation)**.** Explicit regularisation is decomposing a problem into $\mathcal{A} \subseteq \mathcal{B}$, and splitting the excess risk over this, into

$$\underbrace{r(A) - r(a^{**})}_{\text{excess risk}} = \underbrace{r(A) - r(a^*)}_{\text{estimation error}} + \underbrace{r(a^*) - r(a^{**})}_{\text{approx error}}$$

The word "regularisation" indicates we have constrained optimisation, as the estimation error decomposes into constrained optimisations over $\mathcal{A}$.

Proposition 11.6 (Generalization-Optimisation decomposition)**.** *For any $A \in \mathcal{B}$ we have*

$$\mathbb{E}r(A) - r(a^{**}) \le \mathbb{E}\left[r(A) - R(A)\right] + \mathbb{E}\left[R(A) - R(A^{**})\right]$$

By decomposing with a third term that is ≤ 0

Remark 11.7. Generalisation error: can be minimised by choosing a stupid algorithm that returns the same thing regardless of the data.

Definition 11.8 (Algorithmic stability). Setup: Let $A \in \mathcal{B}$ be a given algorithm, which is a function of the RVs $Z_1, ..., Z_n$. let $\tilde{Z}_1, ..., \tilde{Z}_n$ be resampled RVs from the same dist. Let $\tilde{A}(i)$ be the output of A given the perturbed dataset $\{Z_1, ..., Z_{i-1}, \tilde{Z}_i, Z_{i+1}, ..., Z_n\}$

Proposition 11.9 (Generalisation err bound via algorithmic stability)**.** *Given the algorithmic stability setup, we have*

$$\mathbb{E}\left[r(A) - R(A)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell(A, \tilde{Z}_i) - \ell(\tilde{A}(i), \tilde{Z}_i)\right]$$

And if the function $a \mapsto \ell(a, z)$ is γ-Lipschitz for any $z \in \mathcal{Z}$wrt $\|\cdot\|$,

$$\mathbb{E}\left[r(A) - R(A)\right] \le \frac{\gamma}{n}\sum_{i=1}^{n}\mathbb{E}\|A - \tilde{A}(i)\|$$

11.2 Stability for SGD

generalisation error SGD for generic, unconstrained ERM, with multiple passes.

As in 11.8, let $\tilde{W}_t(i)$ be the perturbed version of W_t with dataset $\{Z_1, ..., Z_{i-1}, \tilde{Z}_i, Z_{i+1}, ..., Z_n\}$, i.e.

$$W_t = f_t(Z_1, ..., Z_{i-1}, Z_i, Z_{i+1}, ..., Z_n, I_2, ..., I_t)$$
$$\tilde{W}_t = f_t(Z_1, ..., Z_{i-1}, \tilde{Z}_i, Z_{i+1}, ..., Z_n, I_2, ..., I_t)$$

Lemma 11.10 (Generalisation error bound for convex, Lipschitz and smooth losses)**.** *If, for all $z \in \mathcal{Z}$ $\mathbb{R}^d \ni w \mapsto \ell(w, z)$ is convex, γ-Lipschitz and β-smooth wrt $\|\cdot\|_2$, then SGD with $\eta_s \equiv \eta$ st $\eta\beta \leq 2$ yields, for $t \geq 1$:*

$$\mathbb{E}\|W_t - \tilde{W}_t(i)\|_2 \leq \frac{2\eta\gamma}{n}(t-1),$$

so

$$\mathbb{E}\left[r(W_t) - R(W_t)\right] \leq \frac{2\eta\gamma^2}{n}(t-1)$$

N.B such a loss function might be logistic?

Definition 11.11 (Implicit regularisation)**.** 11.10 shows us that the stability term increases linearly with time, whilst the optimisation error will decrease with time, so if we choose t to minimise the sum of these two bounds this is early stopping, which is a form of implicit/algorithmic regularisation.

Proposition 11.12 (Non-expansivity of gradient update)**.** *If f is β-smooth, convex and $\eta\beta \leq 2$ with $\eta > 0$, then $\forall x, y \in \mathbb{R}$*

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\|_2 \leq \|x - y\|_2$$

12 High dimensional statistics,Gaussian Complexity

Remark 12.1. **Synopsis:** estimating parameters, where we have v. little data. Under sparsity assumption, it is possible to bound error of estimator with Gaussian complexity, if also have column normalisation, then have a w.p bound

Definition 12.2 (Parameter estimation problem)**.** given some training data, and an assumed parametric model for the distribution, infer the parameters of the distribution the data is drawn from. We consider the high-dimensional statistics setting, where the number of parameters d is larger than the number of samples available, n.

Setup:

- data pairs $(x_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, where the x_i are deterministic
- the observations Y_i satisfy $Y_i = \langle x_i, w^* \rangle + \sigma\xi_i$ for some unknown $w^* \in \mathbb{R}^d$, where $\xi_i \sim N(0, 1)$ is IID normal, and $\sigma > 0$ is the standard deviation.
- in matrix form, $Y = \boldsymbol{x}w^* + \sigma\boldsymbol{\xi}$, where $Y \in \mathbb{R}^n$, $\boldsymbol{x} \in \mathbb{R}^{n\times d}$ is the matrix with i-th row $= x_i$.
- We want an estimator W, that is a function of $\boldsymbol{x}, Y$ that minimises $\|W - w^*\|_2$ in expectation/high probability.

Remark 12.3 (Difference between estimation and prediction). As an example, for square loss w/ linear predictors, so $Y = \langle X, w^* \rangle + \sigma\xi$, the prediction error of a given W is

$$\mathbb{E}\left[(\langle X, W \rangle - Y^2 \mid W\right] - \mathbb{E}(\langle X, w^* \rangle - Y)^2] = (W - w^*)^\top \mathbb{E}[XX^\top](W - w^*)$$

but the estimation error is

$$\|W - w^*\|_2^2 == (W - w^*)^\top (W - w^*)$$

So the estimation error might be very large, but aligned with a v. small eigenvalue of $\mathbb{E}[XX^\top]$, so the prediction error is small (or vice versa)

Definition 12.4 (Assumptions on w^*)**.**

- **Sparsity**: the number of non-zero components of w^* is bounded, so $\|w^*\|_0 := \sum_{i=1}^d \mathbf{1}_{|w_i^*|>0} \leq k$ (not really a norm!)

- **Low-rank:** if we can rearrange w^* as a matrix in $\mathbb{R}^{d_1 \times d_2}$where $d_1 \times d_2 = d$, then $\text{Rank}(w^*) \leq k$

Definition 12.5 (Assumptions on $\boldsymbol{x}$).

- **Restricted eigenvalues condition**: $\exists \alpha > 0$ st

$$\alpha \|w\|_2^2 \leq \frac{1}{2n} \|\boldsymbol{x} w\|_2^2$$

for any $w \in \mathbb{R}^d$ with $\|w\|_0 \leq 2k$

- **Restricted isometry property (not used)**: with parameter $2k$ if $\exists \delta \in (0,1)$ st

$$(1-\delta)\|w\|_2^2 \leq \frac{1}{2n} \|\boldsymbol{x} w\|_2^2 \leq (1+\delta)\|w\|_2^2$$

for any $w \in \mathbb{R}^d$ with $\|w\|_0 \leq 2k$

- **Column normalization:** the $\ell_2$norm of each column of $\boldsymbol{x}$ is less than $\sqrt{n}$ - or equivalently each diagonal entry of the second moment matrix is less than 1, as

$$\boldsymbol{c}_{jj} = \frac{1}{n} \sum_{i=1}^n x_{i,j}^2 \leq 1$$

Remark 12.6 (On the restricted eigenvalues condition). The empirical/sample second moment matrix is

$$\boldsymbol{c} := \frac{\boldsymbol{x}^\top \boldsymbol{x}}{n} = \frac{1}{2} \sum_{i=1}^n x_i x_i^\top$$

This is symmetric and real-valued, so its eigenvalues are real. The restricted eigenvalue condition is thus equivalent to

$$\frac{w^\top \boldsymbol{c} w}{w^\top w} \geq 2\alpha \quad \text{for any } w \in \mathbb{R}^d \backslash \{0\} \text{ with } \|w\|_0 \leq 2k$$

Note we can also define the restricted isometry property

Definition 12.7 (sparse estimator).

$$W^0 := \underset{w : \|w\|_0 \leq k}{\arg\min} \frac{1}{2n} \|\boldsymbol{x} w - Y\|_2^2$$

Theorem 12.8 (sparse estimator bound). *If the restricted eigenvalues condition holds, then*

$$\|W^0 - w^*\|_2 \leq \sqrt{2} \frac{\sigma \sqrt{k}}{\alpha} \frac{\|\boldsymbol{x}^\top \xi\|_\infty}{n}$$

Definition 12.9 (Gaussian complexity). : $\xi_i, ..., \xi_n$ are IID $N(0,1)$

$$\text{Gauss}(\mathcal{T}) := \mathbb{E} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \xi_i t_i$$

Proposition 12.10 (Gaussian bounds for sparse estimator). *Given the class of functions* $\mathcal{A}_1 := \big\{ \mathbb{R}^d \ni x \mapsto \langle u, x \rangle \in \mathbb{R} : u \in \mathbb{R}^d, \|u\|$

$$\mathbb{E} \frac{\|\boldsymbol{x}^\top \xi\|_\infty}{n} = \text{Gauss}(\mathcal{A}_1 \circ \{x_1, ..., x_n\})$$

So by theorem 12.8,

$$\mathbb{E}\|W^0 - w^*\|_2 \leq \sqrt{2} \frac{\sigma \sqrt{k}}{\alpha} \text{Gauss}(\mathcal{A}_1 \circ \{x_1, ..., x_n\})$$

Proposition 12.11 (Column normalisation bound in probability). *If the column normalization assumption holds, then*

$$\mathbb{P}\left(\frac{\|\boldsymbol{x}^\top \xi\|_\infty}{\sqrt{b}} \geq \varepsilon \right) \leq 2d \exp(-\varepsilon^2/2)$$

And if <u>sparsity</u> ($\|w^*\|_0 \leq k$) and <u>restricted eigenvalues</u> conditions also hold, then for any $\tau > 2$ we have

$$\mathbb{P}\left(\|W^0 - w^*\|_2 < \frac{\sigma}{\alpha} \sqrt{\frac{2k\tau \log d}{n}} \right) \geq 1 - \frac{2}{d^{(\tau/2-1)}}$$

13 Lasso estimator, Proximal Gradient Methods

Remark 13.1. We have consider the statistical performance of the W^0 estimator (12.7), but it is difficult to compute, in part because the ℓ_0 pseudo-norm makes the optimisation problem non-convex.

Definition 13.2 (Lasso estimator). For a given empirical loss function $R : \mathbb{R}^d \to \mathbb{R}_+$, the Lasso estimator is

$$W^{p1} := \arg\min_{w \in \mathbb{R}^d} R(w) + \lambda \|w\|_1,$$

specifically with the choice $R(w) = \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2$.

Note that we could have specified this as a constrained optimisation without the penalty $\lambda\|w\|_1$term (which would have equivalent power), but the penalised version is more useful in practise as it is more robust against misspecification.

Assumption 13.3 (Restricted strong convexity). *Let $R : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and let $S = supp(w^*) := \{i \in [d] : w_i^* \neq 0\}$ be the support of w^* (so $|supp(w)| = \|w\|_0$, and S^C its complement. Define the cone set:*

$$\mathcal{C} := \left\{ w \in \mathbb{R}^d : \|w_{S^c}\|_1 \leq 3\|w_S\|_1 \right\}$$

Assume $\exists \alpha > 0$ st

$$R(w^* + w) \geq R(w^*) + \langle \nabla R(w^*), w \rangle + \alpha \|w\|_2^2 \quad \textit{for any } w \in \mathcal{C}$$

Remark 13.4 (Connection with strong convexity). This is weaker than strong convexity of R on $\mathcal{C}$, as we require a quadratic lower bound only at w^*, not all of $\mathcal{C}$.

Remark 13.5 (RSC for the ℓ_2 norm). If $R(w) = \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2$, then $\nabla R(w) = \frac{1}{n}\boldsymbol{x}^\top(\boldsymbol{x}w - Y)$, and the assumption simplifies to

$$\frac{1}{2n}\|\boldsymbol{x}w\|_2^2 \geq \alpha\|w\|_2^2 \quad \text{for any } w \in \mathcal{C}$$

Which is very similar to the restricted eigenvalues condition, with the exception of the set w must be in - here $\mathcal{C}$ uses the $\|\cdot\|_1$ norm, but there the $\|\cdot\|_0$ norm. $\mathcal{C}$ is not convex.

Theorem 13.6 (Bound for restricted strong convexity). *If the restricted strong convexity bound holds, and $\lambda \geq 2\|\nabla R(w^*)\|_\infty$, then*

$$\|W^{p1} - w^*\|_2 \leq \frac{3}{2}\frac{\lambda\sqrt{\|w^*\|_0}}{\alpha}$$

Remark 13.7 (Extending 12 to the W^{p1} estimator). If we use the ℓ^2norm, we have that

$$\|\nabla R(w^*)\|_\infty = \sigma\frac{\|\boldsymbol{x}^\top\xi\|_\infty}{n}$$

So the result of 13.6 with $\lambda = 2\|\nabla R(w^*)\|_\infty$ is the same as 12.8 with $k = \|w^*\|_0$ (up to a fact of 3 vs $\sqrt{2}$), so the bounds in expectation and probability in 12 apply to W^{p1} as well - 12.10 and 12.11, with their extra assumptions.

Proposition 13.8 (Sufficient condition for restricted strong convexity). *Define $\|M\| := \max_{i,j}|M_{i,j}|$. If*

$$\|\boldsymbol{c} - I\| \leq \frac{1}{32\|w^*\|_0}$$

Then the restricted strong convexity assumption holds for $R(w) = \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2$ with $\alpha = 1/4$, as for any $w \in \mathcal{C}$, $\frac{1}{2n}\|\boldsymbol{x}w\|_2^2 \geq \|w\|_2^2/4$.

Proposition 13.9 (High probability bound for Rademacher ensembles). *Let $X \in \mathbb{R}^{n\times d}$ be a random matrix with indep. Rademacher components - i.e. $\mathbb{P}(X_{ij} = 1) = \mathbb{P}(X_{ij} = -1) = 1/2$. If $n \geq 2048\tau(\|w^*\|_0)^2\log d$, then for any $\tau \geq 0$*

$$\mathbb{P}\left(\left\|\frac{X^\top X}{n} - I\right\| < \frac{1}{32\|w^*\|_0}\right) \geq 1 - \frac{2}{d^{\tau-2}}$$

13.1 Proximal Gradient Descent

Remark 13.10. So we want to solve

$$\arg\min_{w\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2 + \lambda\|w\|_1.$$

Let $H(w) := \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2 + \lambda\|w\|_1$, which is a convex function, but not strictly convex when $n < d$, so the minima are not unique. If $\boldsymbol{x}$ is drawn from a continuous dist, there may be a unique sol.

The general form of this problem (for a fixed realisation of Y) is

$$\arg\min_{x\in\mathbb{R}^d} h(x);\ h(x) := f(x) + g(x)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and $g : \mathbb{R}^d \to \mathbb{R}$.

Remark 13.11 (Setting up the proximal method). Like in (S)GD, where we move the point that guarantees the local decrease given by the quadratic upper bound from smoothness, we do the same to minimise $f + g$. The smoothness bound on f is

$$f(y) + g(y) \leq g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2 \quad \text{for any } y \in \mathbb{R}^d$$

And the minimisation of this upper bound is

$$\begin{aligned}\arg\min_{y\in\mathbb{R}^d} \{\text{bound...}\} &= \arg\min_{y\in\mathbb{R}^d} \left\{\frac{1}{\beta}g(y) + \frac{1}{2}\left\|y - \left(x - \frac{1}{\beta}\nabla f(x)\right)\right\|_2^2\right\} \\ &=: \mathrm{Prox}_{g/\beta}\left(x - \frac{1}{\beta}\nabla f(x)\right)\end{aligned}$$

Definition 13.12 (Proximal operator)**.** for a function $\kappa : \mathbb{R}^d \to \mathbb{R}$

$$\mathrm{Prox}_\kappa(x) := \arg\min_{y\in\mathbb{R}^d} \left\{\kappa(y) + \frac{1}{2}\|y - x\|_2^2\right\}$$

Algorithm 5 Proximal gradient method

```
1: Input: $x_1, \{\eta_s\}_{s\geq 1}$, stopping time $t$
2: for $s = 1, ..., t$ do
3:     $x_{s+1} \leftarrow \mathrm{Prox}_{\eta_s g}(x_s - \eta_s \nabla f(x_s))$
4: end for
```

Remark 13.13 (on PGD). If $g = 0$, then we get normal gradient descent, and if $g(x) = \infty\mathbf{1}_{x\notin\mathcal{C}}$, then we get projected gradient descent.

N.B. this can be accelerated like GD (see https://blogs.princeton.edu/imabandit/2013/04/01/acceleratedgradientdescent/)

Also, if g is decomposable as $g(x) = \sum_{i=1}^d g_i(x_i)$, then the Prox operator simplifies to solving d 1D convex optimisation problems - which si true for Lasso, as $g(x) = \lambda\sum_{i=1}^d |x_i|$.

Theorem 13.14 (PGD convergence)**.** *If g is convex and β-smooth, and g is just convex, and $\|x_1 - x^*\| \leq b$ then PGD to minimise $h = f + g$ with $\eta_s = 1/\beta$ satisfies*

$$h(x_t) - h(x^*) \leq \frac{\beta b^2}{2(t-1)}$$

Proposition 13.15 (for proof of PGD convergence)**.** *for any $x, y \in \mathbb{R}^d$ we have*

$$h(y) - h(\rho(x)) \geq \frac{\beta}{2}\|\rho(x) - x\|_2^2 + \beta\langle x - y, \rho(x) - x\rangle$$

where $\rho(x) := \mathrm{Prox}_{g/\beta}(x - \frac{1}{\beta}\nabla f(x))$

13.2 Computing the Lasso estimator

Definition 13.16 (soft-thresholding operator). for any $\theta > 0, w \in \mathbb{R}$:

$$\iota(w;\theta) := \mathrm{Prox}_{\theta|\cdot|}(w) = \arg\min_{y\in\mathbb{R}} \left\{\theta|y| + \frac{1}{2}(y-w)^2\right\} = \mathrm{sign}(w)\max\{|w|-\theta, 0\} = \begin{cases} w-\theta & w>0 \\ 0 & w\in[-\theta,\theta] \\ w+\theta & w<-\theta \end{cases}$$

$$W_{s+1} \leftarrow \mathrm{Prox}_{\eta_s g}(W_s - \eta_s \nabla R(W_s)) \equiv \iota(W_s - \frac{\eta_s}{n}\boldsymbol{x}^\top(\boldsymbol{x}W - Y); \lambda\eta_s)$$

Algorithm 6 Iterative Shrinkage-Thresholding Algorithm (ISTA)

```
Setup: $f(w) = R(w) = \frac{1}{2n}\|\boldsymbol{x}w - Y\|_2^2$, $g(w) = \lambda\|w\|_1$
Input: $W_1, \{\eta_s\}_{s\geq 1}$, stopping time $t$
for $s = 1, ..., t$ do
    $W_{s+1} \leftarrow \mathrm{Prox}_{\eta_s g}(W_s - \eta_s \nabla R(W_s)) \equiv \iota(W_s - \frac{\eta_s}{n}\boldsymbol{x}^\top(\boldsymbol{x}W - Y); \lambda\eta_s)$
end for
```

Corollary 13.17 (Applying PGD bounds to ISTA). *Since ISTA is a form of PGD, we can simply apply 13.14 with $\eta = 1/\beta$, where β is the largest eigenvalue of $\boldsymbol{c}$ to get*

$$R(W_t) + \lambda\|W_t\|_1 - (R(W^{p1}) + \lambda\|W^{p1}\|_1) \leq \beta\frac{\|W_1 - W^{p1}\|_2^2}{2(t-1)}$$

Remark 13.18 (applying ISTA). The ultimate goal is bound the deviation $\|W_t - w^*\|_2$, so we would like to use

$$\|W_t - w^*\|_2 \leq \underbrace{\|W_t - W^{p1}\|_2}_{\text{optimisation error}} + \underbrace{\|W^{p1} - w^*\|_2}_{\text{statistics error}}$$

which will combine the optimisation bounds we have just found with the statistical error bound in 13.6.

However, 13.17 bounds the objective function $h(W_t)$ not W_t itself, so we need other results to prove that under restricted strong convexity the iterates W_t also converge (see extra references)

14 Least Squares Regression, Implicit Bias & Implicit Regularization

Summary 14.1. New assumption, that the unknown parameter weights are in the span of the dataset, and instead of adding an explicit regularisation term (Lasso) we optimise the minimisation algorithm of $R(W_t)$ to find us a stopping time to make the estimation error small.

Using the stanard least squares regression setting.

Assumption 14.2 (data span assumption). $\exists\omega \in \mathbb{R}^n$ *st*

$$w^* = \boldsymbol{x}^\top\omega = \sum_{i=1}^n \omega_i x_i$$

Definition 14.3 (Decompositions of empirical second moment). As before, $\boldsymbol{c} := \boldsymbol{x}^\top\boldsymbol{x}/n$ is the empirical second moment matrix. Since it is symmetric & positive semi-definite it has the following orthonormal eigendecomposition

$$\boldsymbol{c} = \boldsymbol{u}\boldsymbol{\mu}\boldsymbol{u}^\top,$$

where $\boldsymbol{u} = [u_1, ..., u_d]$ is an orthonormal basis of eigenvectors of $\boldsymbol{c}$, and $\boldsymbol{\mu}$ contains the corresponding real-valued eigenvalues.

$$\boldsymbol{\mu} := \mathrm{diag}(\mu_1 ..., \mu_r, 0, ..., 0)$$

where r is the rank of the matrix, and the rest of the diagonal is 0. Thus, we can also write

$$\boldsymbol{c} = \boldsymbol{u}_{1:r}\boldsymbol{\mu}_{1:r}\boldsymbol{u}_{1:r}^\top$$

as the removed columns/rows/entries don't contribute.

$\boldsymbol{\pi} := \boldsymbol{u}_{1:r}\boldsymbol{u}_{1:r}^\top$ is the orthogonal projection onto the range of $\boldsymbol{c}$, and the pseudoinverse is $\boldsymbol{c}^+ := \boldsymbol{u}\boldsymbol{\mu}^+\boldsymbol{u}$, where $\boldsymbol{\mu}^+ := \mathrm{diag}(1/\mu_1, ..., 1/\mu_r, 0, ..., 0)$.

14.1 Least squares without Regularisation

The empirical risk is $R(w) = \frac{1}{n}(\boldsymbol{x}w - Y)^2$, so the gradient is $\nabla R(w) = \frac{2}{n}\boldsymbol{x}^\top(\boldsymbol{x}w - Y)$. First order optimality, to find a local minima W^* gives us

$$\boldsymbol{c}W^* = \frac{\boldsymbol{x}^\top Y}{n}.$$

If $\boldsymbol{c}$ is invertible, then ERM has a unique solution

$$W^* = \boldsymbol{c}^{-1}\frac{\boldsymbol{x}^\top Y}{n} = w^* + \sigma \boldsymbol{c}^{-1}\frac{\boldsymbol{x}^\top \xi}{n},$$

using the fact that $Y = \boldsymbol{x}w^* + \sigma\xi$, with some added noise $\sigma\xi$.

If $\boldsymbol{c}$ is not invertible, there are infinitely many solutions, and the solution with smallest Euclidean norm is given by [not proved]:

$$W^*_{l.s.} = \boldsymbol{c}^+\frac{\boldsymbol{x}^\top Y}{n} = \arg\min\left\{\|w\|_2 : \boldsymbol{c}w = \frac{\boldsymbol{x}^\top Y}{n}\right\} = \boldsymbol{\pi}w^* + \sigma\boldsymbol{c}^+\frac{\boldsymbol{x}^\top \xi}{n},$$

since $\boldsymbol{c}^\top\boldsymbol{c} = \boldsymbol{\pi}$.

With regularisation is derived the same way.

14.2 Gradient descent for Least squares

Storing $\boldsymbol{c}$ costs $O(nd^2)$ space, and inversion costs either $O(d^3)$ exactly, or $\tilde{O}(d^2 \log \frac{1}{\varepsilon})$ for an aproximate inverse. Instead of this, we show that if $\mu_r \geq c$ (a universal constant), then gradient descent will solve it up to $1/n$ fast rate in $\tilde{O}(nd)$ time. The gradient descent step is

$$W_{t+1} = W_t - \frac{\eta}{2}\nabla R(W_t) = (I - \eta\boldsymbol{c})W_t + \eta\frac{\boldsymbol{x}^\top Y}{n},$$

a single iteration of requires $O(nd)$ space and time, as $\boldsymbol{c}W_t = \frac{1}{n}\boldsymbol{x}^\top(\boldsymbol{x}W_t)$, and multiplication in this bracketing requries nd operations for each step.

Assuming $W_0 = 0$, then

$$W_t = \left(\sum_{k=0}^{t-1}(I - \eta\boldsymbol{c})^k\right)\eta\frac{\boldsymbol{x}^\top Y}{n} = \mathrm{Inv}_t(\eta\boldsymbol{c})\eta\frac{\boldsymbol{x}^\top Y}{n}, \text{ where } \mathrm{Inv}_t(M) := \sum_{k=0}^{t-1}(I - M)^k$$

Since $Y = \boldsymbol{x}^\top w^* + \sigma\xi$,

$$W_t = \underbrace{\mathrm{Inv}_t(\eta\boldsymbol{c})\eta\boldsymbol{c}w^*}_{\mathbb{E}W_t} + \underbrace{\sigma\,\mathrm{Inv}_t(\eta\boldsymbol{c})\eta\frac{\boldsymbol{x}^\top\xi}{n}}_{W_t - \mathbb{E}W_t}$$

Definition 14.4 (Shrinkage matrix)**.** is

$$\boldsymbol{s} := I - \eta\boldsymbol{\mu} = \mathrm{diag}(1 - \eta\mu_1, ..., 1 - \eta\mu_r, \underbrace{1, ..., 1}_{d-r \text{ times}})$$

Definition 14.5 (Operator norm)**.** for a (possibly non-square) matrix $\boldsymbol{m}$:

$$\|\boldsymbol{m}\| := \sqrt{\mu_1(\boldsymbol{m}^\top\boldsymbol{m})},$$

where $\mu_d(\boldsymbol{m}^\top\boldsymbol{m}) \leq \cdots \leq \mu_1(\boldsymbol{m}^\top\boldsymbol{m})$ are the real eigenvalues of the symmetric matrix $\boldsymbol{m}^\top\boldsymbol{m}$. If $\boldsymbol{m}$ is symmetric, then $\|\boldsymbol{m}\| = \max\{\mu_1(\boldsymbol{m}), \mu_d(\boldsymbol{m})\}$.

Note also that

$$\|\boldsymbol{m}\tilde{\boldsymbol{m}}\| \leq \|\boldsymbol{m}\|\|\tilde{\boldsymbol{m}}\|$$

Proposition 14.6 (t-Inverses of second order moments)**.**

$$\mathrm{Inv}_t(\eta\boldsymbol{c}) = (I - \boldsymbol{u}\boldsymbol{s}^t\boldsymbol{u}^\top)(\eta\boldsymbol{c})^+ + t(I - \boldsymbol{\pi}) = \sum_{i-1}^{r}\frac{1 - (1 - \eta\mu_i)^t}{\eta\mu_i}u_iu_i^\top + t(I - \boldsymbol{\pi})$$

$$\mathrm{Inv}_t(\eta\boldsymbol{c})\eta\boldsymbol{c} = (I - \boldsymbol{u}\boldsymbol{s}^t\boldsymbol{u}^\top) = \sum_{i-1}^{r}\frac{1 - (1 - \eta\mu_i)^t}{\eta\mu_i}u_iu_i^\top$$

Which we can use in the decomposition of W_t.

Proposition 14.7. *If $\eta \leq \frac{1}{\mu_1}$ then*

$$\lim_{t \to \infty} (I - \boldsymbol{u}\boldsymbol{s}^t \boldsymbol{u}^\top) = \boldsymbol{\pi},$$

so the limit of W_t as $t \to \infty$ is $W^*_{l.s.}$, and more importantly

$$\|W_t - W^*_{l.s.}\|_2 \leq (1 - \eta\mu_r)^t \|w^*\|_2 + \frac{\sigma(1-\eta\mu_1)^t}{\sqrt{n}\mu_r} \left\| \frac{\boldsymbol{x}^\top \xi}{\sqrt{n}} \right\|_2$$

NO PROOF OF IMPORTANT BIT??????

Remark 14.8. It's not surprising that gradient descent finds the min-ℓ_2 norm solution, as it essentially uses the $\ell_2$norm in each time step. This is called implicit bias

14.3 Implicit regularization

Theorem 14.9. *As a simple decomposition,*

$$\|W_t - w^*\|_2 \leq \underbrace{\|\mathbb{E}W_t - \boldsymbol{\pi} w^*\|_2}_{\textit{bias error}} + \underbrace{\|W_t - \mathbb{E}W_t\|_2}_{\textit{concentration error}} + \underbrace{\|w^* - \boldsymbol{\pi} w^*\|_2}_{\textit{approx error}}.$$

If $\eta \leq \frac{1}{\mu_1}$, then $\|\mathbb{E}W_t - \boldsymbol{\pi} w^*\|_2 \leq (1-\eta\mu_r)^t \|w^*\|_2$ and $\|W_t - \mathbb{E}W_t\|_2 \leq \frac{\sigma}{\sqrt{n}} \frac{1-(1-\eta\mu_1)^t}{\mu_r} \frac{\|\boldsymbol{x}^\top \xi\|_2}{\sqrt{n}}$.

Further, for $c \in (0,1)$,if t^* st $t^* \geq \frac{1}{\log(1/(1-\eta\mu_r))} \log\left(\frac{\|w^*\|_2}{\sigma}\sqrt{n}/\tilde{c}\right)$, where $\tilde{c} := \frac{1}{\mu_r}\sqrt{\sum_{i=1}^r \mu_i + c\sum_{i=1}^r \mu_i^2/\mu_1}$, then

$$\mathbb{P}\left(\|W_{t^*} - w^*\|_2 \leq 2\sigma\frac{\tilde{c}}{\sqrt{n}} + \|w^* - \boldsymbol{\pi} w^*\|_2\right) \geq 1 - \delta,$$

where $\delta = \exp\left(-c^2/8 \times \sum_{i=1}^r (\mu_i/\mu_1)^2\right)$

Remark 14.10. If the eigenvalues are upper and lower bounded by constants indep of n, d,etc., and if the signal to noise ratio $\|w^*\|_2/\sigma$ is upper bounded as well, then we get the fast rate (of the square of the $\ell_2$loss) .

14.4 Alternative decomposition

Proposition 14.11 (Bias-variance decompositions)**.** *Assuming $\boldsymbol{\pi} w^* = w^*$ (so no approx error), then*

$$\mathbb{E}\|W_t - w^*\|_2^2 \leq \underbrace{\|\mathbb{E}W_t - w^*\|_2^2}_{\textit{square of bias err}} + \underbrace{\mathbb{E}\|W_t - \mathbb{E}W_t\|_2^2}_{\textit{variance error}} = \underbrace{\|\mathbb{E}W_t - w^*\|_2^2}_{\textit{square of bias err}} + \underbrace{\sum_{i=1}^d \mathrm{Var}(W_{t,i})}_{\textit{variance error}}$$

$$\|W_t - w^*\|_2^2 \leq \underbrace{2\|\mathbb{E}W_t - w^*\|_2^2}_{\textit{square of bias err}} + \underbrace{2\|W_t - \mathbb{E}W_t\|_2^2}_{\textit{concentration error}}$$

15 Stochastic Multi-Armed Bandits: Problem and Algorithms

Definition 15.1 (Online statistical learning)**.** At each time step $t = 1, 2, ..., n$:

- choose an action $A_t \in \mathcal{A}$ out of the admissible set (possibly random)
- a data point $Z_t \in \mathcal{Z}$ is sampled from an unknown distribution.
    - **which setting\/????**
- suffer a loss $\ell(A_t, Z_t)$.
- update the normalised pseudo-regret:

  $$\frac{1}{n}\sum_{t=1}^n r(A_t) - \inf_{a \in \mathcal{A}} r(a),$$

  where r is the standard expected/population risk.

Goal: minimise the pseudo-regret.

Definition 15.2 (Stochastic multi-armed bandit)**.** As above, with

- the action set is finite $\mathcal{A} = \{1, ...k\}$, each action is called an arm
- $Z_t = (Z_{t,1}, ..., Z_{t,k}) \in \mathcal{Z} = [0,1]^k$ is a vector of independently sampled components. $Z_{t,a}$ is sampled from a distribution D_a iid, with mean μ_a.
- Z_t is not revealed to the player, as this is bandits.
- the loss is $\ell(A_t, Z_t) = -Z_{t,A_t}$, so $r(a) = -\mu_a$, and the normalised pseudo-regret is
$$\max_{a \in \mathcal{A}} \mu_a - \frac{1}{n} \sum_{t=1}^{n} \mu_{A_t}$$
- let $a^* \in \arg\max_{a \in \mathcal{A}} \mu_a$,so the (unnormalised) pseudo-regret at time n is
$$R_n := n\mu_{a^*} - \sum_{t=1}^{n} \mu_{A_t}$$
- which we want to minimise, given A_t depends only on $(A_s, \ell(A_s, Z_s))_{s \in [t-1]}$
- let $\Delta_a := \mu_{a^*} - \mu_a$ be the sub-optimality gap of arm a.
- the number of times arm a is pulled up to time t is
$$N_{t,a} := \sum_{s=1}^{t} \mathbf{1}_{A_s = a}$$
- and the sample mean of the rewards from playing a up to time t is
$$M_{t,a} := \frac{1}{N_{t,a}} \sum_{s=1}^{t} Z_{s,a} \mathbf{1}_{A_s = a}$$

Proposition 15.3.

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a N_{n,a}$$

Algorithm 7 Explore-Then-Commit(ε)

```
Input: $\varepsilon \in \mathbb{N}_+$
for $t = 1, ..., \varepsilon k$ do
    set $A_t = 1 + (t-1) \mod k$
end for
for $t = \varepsilon k, ..., n$ do
    set $A_t \in \arg\max_{a \in \mathcal{A}} M_{\varepsilon k, a}$
end for
```

Proposition 15.4 (Pseudo-Regret for Explore-then-Commit)**.** *: for any $\varepsilon \in \mathbb{N}_+$, there is a stochastic multi-armed bandit problem st the expected pseudo risk is*

$$\mathbb{E}R_n = cn + \tilde{c},$$

for constants $c, \tilde{c} \in \mathbb{R}_+$ that are independent of n.

Algorithm 8 Greedy(ε)

```
Input: $\varepsilon \in (0,1)$
for $t = 1, ..., k$ do
    set $A_t = t$
end for
for $t = k+1, ..., n$ do
    set $A_t \in \begin{cases} \in \arg\max_{a \in \mathcal{A}} M_{t-1,a} & \text{w.p. } 1-\varepsilon \\ \sim Unif\{1, ..., k\} & \text{w.p. } \varepsilon \end{cases}$
end for
```

Proposition 15.5 (Pseudo-Regret for Greedy). *: for any $\varepsilon \in \mathbb{N}_+$, there is a stochastic multi-armed bandit problem st the expected pseudo risk is*

$$\mathbb{E}R_n = cn + \tilde{c},$$

for constants $c, \tilde{c} \in \mathbb{R}_+$ that are independent of n

Definition 15.6 (Upper confidence bounds). at time t for algorithm a, the upper confidence bound is

$$U_{t,a} := M_{t,a} + \sqrt{\frac{\varepsilon \log t}{2N_{t,a}}},$$

which estimates the mean reward, optimistically, by calculating a fixed confidence level above it. The estimator decreases as arm a is played more, and increases logarithmically as time passes. ε controls the tradeoff between exploitation of known performance ($M_{t,a}$), and exploring arms we are less certain in (the right term)

Algorithm 9 Upper Confidence Bound (UCB)(ε)

```
Input: $\varepsilon \in \mathbb{R}_+$
for $t = 1, ..., k$ do
    set $A_t = t$
end for
for $t = k + 1, ..., n$ do
    set $A_t \in \arg\max_{a \in \mathcal{A}} U_{t-1,a}$
end for
```

Proposition 15.7 (Pseudo-Regret for UCB). *: for any $\varepsilon \in \mathbb{N}_+$, there is a stochastic multi-armed bandit problem st the expected pseudo risk is*

$$\mathbb{E}R_n \leq \sum_{a \in \mathcal{A}} \frac{2\varepsilon \log n}{\Delta_a} + 2 \sum_{a \in \mathcal{A}} \Delta_a \sum_{t=k}^{n-1} \frac{1}{t^{\varepsilon}}$$

for constants $c, \tilde{c} \in \mathbb{R}_+$ that are independent of n

Proposition 15.8 (UCB proof #1). *for any non-decreasing sequence $s_1 \leq \cdots \leq s_n$ in $\mathbb{R}_+$, and any $a \in \mathcal{A}$*

$$\mathbb{E}N_{n,a} \leq s_n + \sum_{t=k}^{n-1} \mathbb{P}(A_{t+1} = a \mid N_{t,a} \geq s_t)$$

Proposition 15.9 (UCB proof #2). *let $A_{t+1} = \arg\max_{a \in \mathcal{A}} U_{t,a}$, where $U_{t,a} = M_{t,a} + \sqrt{\frac{\log(1/\delta)}{2N_{t,a}}}$, for any any $a \in \mathcal{A}$ st $\Delta_a > 0$,*

$$\mathbb{P}\left(A_{t+1} = a \mid N_{t,a} \geq \frac{2\log(1/\delta)}{\Delta_a^2}\right) \leq 2\delta$$

Proposition 15.10 (UCB proof #3). *for any $a \in \mathcal{A}$:*

$$\mathbb{E}N_{n,a} \leq \frac{2\varepsilon \log n}{\Delta_a^2} + 2 \sum_{t=k}^{n-1} \frac{1}{t^{\varepsilon}}$$

Theorem 15.11 (Distribution-independent bound for UCB). *for any $n \geq k$*

$$\mathbb{E}R_n \leq 2\sqrt{2\varepsilon kn \log n} + 2k \sum_{t=k}^{n-1} \frac{1}{t^{\varepsilon}}$$

16 Minimax lower bounds and Hypothesis testing

Definition 16.1 (Hypothesis testing setup). random variable $X \in \mathcal{X}$, drawn either from a distribution $\mathbb{P}$ (the null hypothesis H_0) or $\mathbb{Q}$, the alternative hypothesis H_1. $f : \mathcal{X} \to \{0, 1\}$ is a test that indicates which hypothesis should be true. There are 2 types of error: **type 1** is $f(X) = 1$ when $X \sim \mathbb{P}$, and **type 2** is $f(X) = 0$ when $X \sim \mathbb{Q}$. $\mathbb{P}, \mathbb{Q}$ have densitites p, q wrt a measure ρ if $\mathbb{P}(E) = \int p(x)\mathbf{1}_E(x)\rho(dx)$, same for $\mathbb{Q}$.

Lemma 16.2 (Neyman Pearson). *for any function $f : \mathcal{X} \to \{0,1\}$ we have*

$$\mathbb{P}(f(X) = 1) + \mathbb{Q}(f(X) = 0) \geq \int \min\{p(x), q(x)\ \rho(dx)$$

and equality is achieved by the likelihood ratio test $f^* := \mathbf{1}_{q \geq p}$

Definition 16.3 (Total Variation (TV) distance). between two distributions $\mathbb{P}, \mathbb{Q}$ defined on the same measurable space with densities p, q wrt ρ

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} := \sup_E |\mathbb{P}(E) - \mathbb{Q}(E)| = \frac{1}{2}\int |p(x) - q(x)|\ \rho(dx) = 1 - \int \min\{p(x), q(x)\ \rho(\mathrm{d}x)$$

so, a lower bound in the Nyman Pearson lemma is an upper bound on the total variation distance.

Definition 16.4 (K-L divergence). between $\mathbb{P}, \mathbb{Q}$ on the same measure space, densities p, q is

$$\texttt{KL}(\mathbb{P}, \mathbb{Q}) := \begin{cases} \int p(x) \log \frac{p(x)}{q(x)}\ \rho(dx) & \text{if } \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{otherwise} \end{cases}$$

where $\mathbb{P} \ll \mathbb{Q}$ means $\mathbb{Q}(E) = 0 \implies \mathbb{P}(E) = 0$ for all measurable E.

Proposition 16.5 (Properties of KL divergence). *1.* **Gibb's inequality***: $KL(\mathbb{P}, \mathbb{Q}) \geq 0$, equality if $\mathbb{P} = \mathbb{Q}$*

2. **Chain rule** *for product distributions: if $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$, and $\mathbb{Q} = \otimes_{i=1}^n \mathbb{Q}_i$, then $KL(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^n KL(\mathbb{P}_i, \mathbb{Q}_i)$*

3. **Pinsker's inequality:** *for any measurable event E,*

$$\mathbb{P}(E) - \mathbb{Q}(E) \leq \sqrt{\frac{1}{2} KL(\mathbb{P}, \mathbb{Q})},$$

which implies that

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \leq \sqrt{\frac{1}{2} KL(\mathbb{P}, \mathbb{Q})},$$

and this obviously also holds if we swap $\mathbb{P}, \mathbb{Q}$, even though the KL divergence isn't symmetric.

Corollary 16.6. *Let $X = \{X_1, ..., X_n\} \in \mathcal{X}^n$ be distributed according to $\mathbb{P}$ or $\mathbb{Q}$ on $\mathcal{X}^n$ (i.e. not necessarily products). for any test function $f : \mathcal{X}^n \to \{0,1\}$ we have*

$$\mathbb{P}(f(X_1, ..., X_n) = 1) + \mathbb{Q}(f(X_1, ..., X_n) = 0) \geq 1 - \sqrt{\frac{1}{2} KL(\mathbb{P}, \mathbb{Q})},$$

and if this is a product dist, then the KL in the square root splits to a sum.

Theorem 16.7 (Distribution-independent lower bound for multi-armed bandit). *Let $n \geq k - 1$, and for any algorithm that chooses $A_1, ..., A_n$, there is a k-armed bandit problem st*

$$\mathbb{E}R_n \geq c\sqrt{(k-1)n},$$

for a universal constant c.

Proposition 16.8. *consider 2 k-armed bandit problems, the first labelled μ, the second ν, so the reward probabilities per arm are $P_{\mu,a}$ and $P_{\nu,a}$, with densities $p_{\mu,a}, p_{\nu,a}$ wrt ρ. For an algorithm $A_1, ..., A_n$, let P_μ, P_ν be the probability that each bandit problem assigns to $(A_1, , Z_{1,A_1}, ..., A_n, Z_{n,A_n})$*

$$KL(P_\mu, P_\nu) = \sum_{a=1}^k KL(P_{\mu,a}, P_{\nu,a}) E_\mu N_{n,a}$$

Note that P_μ, P_ν are not products of the per-arm distributions, as they also include the randomness in the algorithm.

Theorem 16.9 (Fano's inequality). *given $\mathbb{P}_1, ..., \mathbb{P}_m$ are prob measures such that $\mathbb{P}_i \ll \mathbb{P}_j$ for any $i, j \in [m]$, then NOT USUAL VERSION*

$$\inf_f \max_{i \in [m]} \mathbb{P}_i(f(X) \neq i) \geq 1 - \frac{\frac{1}{m^2}\sum_{i,j}^m KL(\mathbb{P}_i, \mathbb{P}_j) + \log 2}{\log(m-1)}$$

which is often applicable to minimax bounds (though worse than Neyman Pearson) using covering numbers to reduce a $\sup_w$ to $\max_{i \in [m]}$