# Probability Inequalities for Sequential Random Variables

Mathematical Analysis of Machine Learning Algorithms
(Chapter 13)

# Sequential Estimation

In sequential estimation problems, we observe a sequence of random variables $Z_t \in \mathcal{Z}$ for $t = 1, 2, \ldots$, where each $Z_t$ may depend on the previous observations $\mathcal{S}_{t-1} = [Z_1, \ldots, Z_{t-1}] \in \mathcal{Z}^{t-1}$.

## Notations

The sigma algebra generated by $\{\mathcal{S}_t\}$ forms a natural filtration $\{\mathcal{F}_t\}$.

We say a sequence $\{\xi_t\}$ is adapted to the filtration $\{\mathcal{F}_t\}$, if each $\xi_t$ is a function of $\mathcal{S}_t$. That is, each $\xi_t$ at time $t$ does not depend on future observations $Z_s$ for $s > t$.

Alternatively one may also say that $\xi_t$ is measurable in $\mathcal{F}_t$.

# Martingale

The sequence

$$\xi'_t = \xi_t - \mathbb{E}[\xi_t|\mathcal{F}_{t-1}], \text{ or equivalently } \xi'_t(\mathcal{S}_t) = \xi_t(\mathcal{S}_t) - \mathbb{E}_{Z_t|\mathcal{S}_{t-1}}\xi_t(\mathcal{S}_t),$$

is referred to as a *martingale difference sequence* with the property

$$\mathbb{E}[\xi'_t|\mathcal{F}_{t-1}] = \mathbb{E}_{Z_t|\mathcal{S}_{t-1}}\xi'_t(\mathcal{S}_t) = 0.$$

The sum of a martingale difference sequence

$$\sum_{s=1}^{t} \xi'_s = \sum_{s=1}^{t} \xi'_s(\mathcal{S}_s)$$

is referred to as a *martingale*, which satisfies (for all $t$):

$$\mathbb{E}\left[\sum_{s=1}^{t} \xi'_s|\mathcal{F}_{t-1}\right] = \sum_{s=1}^{t-1} \xi'_s, \text{ or } \mathbb{E}_{Z_t|\mathcal{S}_{t-1}} \sum_{s=1}^{t} \xi'_s(\mathcal{S}_s) = \sum_{s=1}^{t-1} \xi'_s(\mathcal{S}_s).$$

# General Notation

## Refined Notation

We assume each observation is

$$\mathcal{Z} = \mathcal{Z}^{(x)} \times \mathcal{Z}^{(y)},$$

and each $Z_t \in \mathcal{Z}$ can be written as $Z_t = (Z_t^{(x)}, Z_t^{(y)})$.

We are interested in the conditional expectation with respect to $Z_t^{(y)} | Z_t^{(x)}, \mathcal{S}_{t-1}$, rather than with respect to $Z_t | \mathcal{S}_{t-1}$.

Without causing confusion, we adopt the following shortened notation

$$\mathbb{E}_{Z_t^{(y)}}[\cdot] = \mathbb{E}_{Z_t^{(y)} | Z_t^{(x)}, \mathcal{S}_{t-1}}[\cdot].$$

This formulation is useful in many statistical estimation problems such as regression, where conditional expectation is what we are interested in.

# Martingale Exponential Equality

**Lemma 1 (Martingale Exponential Equality, Lem 13.1 )**

*Consider a sequence of real-valued random (measurable) functions $\xi_1(\mathcal{S}_1), \ldots, \xi_T(\mathcal{S}_T)$. Let $\tau \le T$ be a stopping time so that $\mathbb{1}(t \le \tau)$ is measurable in $\mathcal{S}_t$. We have*

$$\mathbb{E}_{\mathcal{S}_T} \exp\left(\sum_{i=1}^{\tau} \xi_i - \sum_{i=1}^{\tau} \ln \mathbb{E}_{Z_i^{(y)}} e^{\xi_i}\right) = 1.$$

# Proof of Lemma 1 (I/II)

We prove the lemma by induction on $T$. When $T = 0$, the equality is trivial. Assume that the claim holds at $T-1$ for some $T \geq 1$. Now we will prove the equation at time $T$ using the induction hypothesis.

Note that $\tilde{\xi}_i = \xi_i \mathbb{1}(i \leq \tau)$ is measurable in $\mathcal{S}_i$. We have

$$\sum_{i=1}^{\tau} \xi_i - \sum_{i=1}^{\tau} \ln \mathbb{E}_{Z_i^{(y)}} e^{\xi_i} = \sum_{i=1}^{T} \tilde{\xi}_i - \sum_{i=1}^{T} \ln \mathbb{E}_{Z_i^{(y)}} e^{\tilde{\xi}_i}.$$

It follows that

$$\begin{aligned}
&\mathbb{E}_{Z_1,\ldots,Z_T} \exp\left(\sum_{i=1}^{\tau} \xi_i - \sum_{i=1}^{\tau} \ln \mathbb{E}_{Z_i^{(y)}} e^{\xi_i}\right) \\
=&\mathbb{E}_{Z_1,\ldots,Z_T} \exp\left(\sum_{i=1}^{T} \tilde{\xi}_i - \sum_{i=1}^{T} \ln \mathbb{E}_{Z_i^{(y)}} e^{\tilde{\xi}_i}\right) \\
=&\mathbb{E}_{Z_1,\ldots,Z_{n-1},Z_T^{(x)}} \left[\exp\left(\sum_{i=1}^{T-1} \tilde{\xi}_i - \sum_{i=1}^{T-1} \ln \mathbb{E}_{Z_i^{(y)}} e^{\tilde{\xi}_i}\right) \underbrace{\mathbb{E}_{Z_T^{(y)}} \exp(\tilde{\xi}_T - \ln \mathbb{E}_{Z_T^{(y)}} e^{\tilde{\xi}_T})}_{=1}\right]
\end{aligned}$$

# Proof of Lemma 1 (II/II)

$\dots$

$$
\begin{aligned}
=&\mathbb{E}_{Z_1,\dots,Z_{n-1},Z_T^{(x)}}\left[\exp\left(\sum_{i=1}^{T-1}\tilde{\xi}_i-\sum_{i=1}^{T-1}\ln\mathbb{E}_{Z_i^{(y)}}e^{\tilde{\xi}_i}\right)\underbrace{\mathbb{E}_{Z_T^{(y)}}\exp(\tilde{\xi}_T-\ln\mathbb{E}_{Z_T^{(y)}}e^{\tilde{\xi}_T})}_{=1}\right]\\
=&\mathbb{E}_{Z_1,\dots,Z_{T-1}}\exp\left(\sum_{i=1}^{T-1}\tilde{\xi}_i-\sum_{i=1}^{T-1}\ln\mathbb{E}_{Z_i^{(y)}}e^{\tilde{\xi}_i}\right)\\
=&\mathbb{E}_{Z_1,\dots,Z_{T-1}}\exp\left(\sum_{i=1}^{\min(\tau,T-1)}\xi_i-\sum_{i=1}^{\min(\tau,T-1)}\ln\mathbb{E}_{Z_i^{(y)}}e^{\xi_i}\right)=1.
\end{aligned}
$$

Note that the last equation follows from the induction hypothesis, and the fact that $\min(\tau, T-1)$ is a stopping time $\leq T-1$.

# Martingale Exponential Tail Inequality

## Theorem 2 (Thm 13.2)

*Consider a sequence of random functions* $\xi_1(\mathcal{S}_1), \ldots, \xi_t(\mathcal{S}_t), \ldots$*, with filtration* $\{\mathcal{F}_t\}$*. We have for any* $\delta \in (0,1)$ *and* $\lambda > 0$*:*

$$\Pr\left[\exists n > 0 : -\sum_{i=1}^{n} \xi_i \geq \frac{\ln(1/\delta)}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda \xi_i}\right] \leq \delta.$$

*Moreover, consider a sequence of* $\{z_t \in \mathbb{R}\}$ *adapted to* $\{\mathcal{F}_t\}$*, and events* $A_t$ *on* $\mathcal{F}_t$*:*

$$\ln \Pr\left[\exists n > 0 : \sum_{i=1}^{n} \xi_i \leq z_n \,\&\, \mathcal{S}_n \in A_n\right]$$
$$\leq \inf_{\lambda>0} \sup_{n>0} \sup_{\mathcal{S}_n \in A_n} \left[\lambda z_n + \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda \xi_i}\right].$$

# Proof of Theorem 2 (I/II)

We will prove the result for a finite time sequence $\xi_1(\mathcal{S}_1), \ldots, \xi_T(\mathcal{S}_T)$. It implies the desired result by letting $T \to \infty$. Let

$$\xi_\tau(\lambda) = -\sum_{i=1}^{\tau} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda \xi_i} - \lambda \sum_{i=1}^{\tau} \xi_i,$$

where $\tau$ is a stopping time, then we have from Lemma 1: $\mathbb{E}\, e^{\xi_\tau(\lambda)} = 1$. Now for any given sequence of $\tilde{z}_n(\mathcal{S}_n)$ and $A_n$, define the stopping time $\tau$ as either $T$, or the first time step $n$ so that

$$\xi_n(\lambda) \geq -\tilde{z}_n(\mathcal{S}_n) \;\&\; \mathcal{S}_n \in A_n$$

for each sequence $\mathcal{S}_T$. It follows that

$$\begin{aligned}
&\Pr\left(\exists n : \xi_n(\lambda) \geq -\tilde{z}_n(\mathcal{S}_n) \;\&\; \mathcal{S}_n \in A_n\right) \inf_{n>0, \mathcal{S}_n \in A_n} e^{-\tilde{z}_n(\mathcal{S}_n)} \\
\leq &\mathbb{E}\left[e^{\xi_\tau(\lambda) + \tilde{z}_\tau(\mathcal{S}_\tau)} \mathbb{1}(\mathcal{S}_\tau \in A_\tau)\right] \inf_{n>0, \mathcal{S}_n \in A_n} e^{-\tilde{z}_n(\mathcal{S}_n)} \\
\leq &\mathbb{E}\left[e^{\xi_\tau(\lambda) + \tilde{z}_\tau(\mathcal{S}_\tau)} \mathbb{1}(\mathcal{S}_\tau \in A_\tau) e^{-\tilde{z}_\tau(\mathcal{S}_\tau)}\right] \leq \mathbb{E}\, e^{\xi_\tau(\lambda)} = 1.
\end{aligned}$$

# Proof of Theorem 2 (II/II)

Therefore we obtain

$$\ln \Pr\left[\exists n > 0 : -\lambda \sum_{i=1}^{n} \xi_i \geq -\tilde{z}_n(\mathcal{S}_n) + \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda \xi_i} \;\&\; \mathcal{S}_n \in A_n\right]$$
$$\leq \sup_{n>0:\mathcal{S}_n \in A_n} \tilde{z}_n(\mathcal{S}_n).$$

Let $\tilde{z}(\mathcal{S}_n) = \ln \delta$, we obtain the first inequality. Let

$$\tilde{z}_n(\mathcal{S}_n) = \lambda z_n + \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda \xi_i},$$

we obtain the second inequality.

# Remarks

- In the iid case (Theorem 2.5), one can optimizing over $\lambda$ to obtain an inequality in terms of the rate function.
- Theorem 13.2 requires fixed $\lambda$. However, one can take a union bound over $\lambda$ to obtain a result that holds for all $\lambda$: this may lead to an extra $\log n$ factor in the resulting bound.
- The second inequality in Theorem 2 resolves the issue of paying extra $\log$ factor by restricting the optimization over $\lambda$ in a restricted event $A_n$.
- In general, one can obtain sample dependent bounds requiring empirical quantities bounded in $A_n$, and this can be alleviated by taking union bound over $A_n$.

# Martingale Sub-Gaussian Inequality

**Theorem 3 (Martingale Sub-Gaussian Inequality, Thm 13.3)**

*Consider a sequence of random functions* $\xi_1(\mathcal{S}_1), \ldots, \xi_t(\mathcal{S}_t), \ldots$. *Assume each* $\xi_i$ *is sub-Gaussian with respect to* $Z_i^{(y)}$*:*

$$\ln \ \mathbb{E}_{Z_i^{(y)}} e^{\lambda \xi_i} \leq \lambda \mathbb{E}_{Z_i^{(y)}} \xi_i + \frac{\lambda^2 \sigma_i^2}{2}$$

*for some* $\sigma_i$ *that may depend on* $\mathcal{S}_{i-1}$ *and* $Z_i^{(x)}$*. Then for all* $\sigma > 0$*, with probability at least* $1 - \delta$*,*

$$\forall n > 0 : \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \left( \sigma + \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma} \right) \sqrt{\frac{\ln(1/\delta)}{2}}.$$

# Data Independent Sub-Gaussian Bound

Since we allow $\sigma_i$ to be data dependent, we cannot in general choose $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. However, if $\sigma_i$ does not depend on data, then we can further optimize $\sigma$ for specific time horizon $n$.

**Theorem 4 (Azuma's Inequality, Thm 13.4)**

*Consider a sequence of random functions $\xi_1(\mathcal{S}_1), \ldots, \xi_n(\mathcal{S}_n)$ with a fixed number $n > 0$. If for each $i$: $\sup \xi_i - \inf \xi_i \leq M_i$ for some constant $M_i$, then with probability at least $1 - \delta$,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \frac{1}{n}\sum_{i=1}^{n} \xi_i + \sqrt{\frac{\sum_{i=1}^n M_i^2 \ \ln(1/\delta)}{2n^2}}.$$

# Example: Data Dependent sub-Gaussian Inequality

We now consider the situation $\sigma_i$ is data dependent in Theorem 3. Using the technique of Chapter 8, we can obtain the following data-dependent bound.

## Proposition 5

*Under the assumptions of Theorem 3. Given any* $c_0 > 0$ *and* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*:*

$$\forall n > 0 : \quad \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \sqrt{4 \left( c_0 + \sum_{i=1}^{n} \sigma_i^2 \right) \ln \frac{(\hat{\ell} + 1)^2}{\delta}},$$

*where* $\hat{\ell} = \lfloor 1 + \log_2(1 + \sum_{i=1}^{n} \sigma_i^2 / c_0) \rfloor$.

# Proof of Proposition 5

Consider the sequence of numbers $2^\ell c_0$ ($\ell = 1, \ldots$). For each $\ell$, we consider the event $\sum_{i=1}^n \sigma_i^2 \leq 2^\ell c_0$, and let $\sigma = \sqrt{2^\ell c_0}$ in Theorem 3. It follows that with probability at least $1 - \delta/(\ell+1)^2$,

$$\forall n > 0 : \sum_{i=1}^n \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^n \xi_i + \sqrt{2^{\ell+1} c_0 \ln \frac{(\ell+1)^2}{\delta}} \text{ or } \sum_{i=1}^n \sigma_i^2 > 2^\ell c_0.$$

Taking union bound, the above inequality holds for all $\ell \geq 1$ with probability at least $1 - \delta$.
Now let $\hat{\ell} = \lfloor 1 + \log_2(1 + \sum_{i=1}^n \sigma_i^2 / c_0) \rfloor$, we know that the following inequalities hold

$$\sum_{i=1}^n \sigma_i^2 \leq 2^{\hat{\ell}} c_0, \qquad 2^{\hat{\ell}+1} c_0 \leq 4 \left( c_0 + \sum_{i=1}^n \sigma_i^2 \right).$$

Therefore we obtain the desired bound.

# Multiplicative Chernoff Bound

## Theorem 6 (Thm 13.5)

*Consider a sequence of random functions* $\xi_1(\mathcal{S}_1), \ldots, \xi_t(\mathcal{S}_t), \ldots$ *such that* $\xi_i \in [0, 1]$ *for all i. We have for* $\lambda > 0$*, with probability at least* $1 - \delta$*:*

$$\forall n > 0 : \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \frac{\lambda}{1 - e^{-\lambda}} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \frac{\ln(1/\delta)}{(1 - e^{-\lambda})\, n}.$$

*Similarly, for* $\lambda > 0$*, with probability at least* $1 - \delta$*:*

$$\forall n > 0 : \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i > \frac{\lambda}{e^{\lambda} - 1} \frac{1}{n} \sum_{i=1}^{n} \xi_i - \frac{\ln(1/\delta)}{(e^{\lambda} - 1)\, n}.$$

We note that similar to Theorem 3, the result is with fixed $\lambda$. However similar to Proposition 5, we can take union bound over a range of $\lambda$ values to obtain a bound that allow $\lambda$ to be data dependent.

# Result used in the Proof of Theorem 6

## Lemma 7 (Lem 2.15 )

*Consider a random variable* $X \in [0,1]$ *and* $\mathbb{E}X = \mu$*. We have the following inequality:*

$$\ln \mathbb{E} e^{\lambda X} \leq \ln[(1-\mu)e^0 + \mu e^\lambda] \leq \lambda\mu + \lambda^2/8.$$

# Proof of Theorem 6

We obtain from Lemma 7 and Theorem 2 that with probability at least $1-\delta$

$$
\begin{aligned}
-\sum_{i=1}^{n} \xi_i <& \frac{\ln(1/\delta)}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{n} \ln(1+(e^{-\lambda}-1)\mathbb{E}_{Z_i^{(y)}}\xi_i) \\
\leq& \frac{\ln(1/\delta)}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{n} (e^{-\lambda}-1)\mathbb{E}_{Z_i^{(y)}}\xi_i.
\end{aligned}
$$

This implies the first bound. The second bound can be proved similarly.

# Freedman's Inequality

## Theorem 8 (Freedman's Inequality, Thm 13.6)

*Consider a sequence of random functions* $\xi_1(\mathcal{S}_1), \ldots, \xi_n(\mathcal{S}_n)$. *Assume that* $\xi_i \geq \mathbb{E}_{Z_i^{(y)}} \xi_i - b$ *for some constant* $b > 0$. *Then for any* $\lambda \in (0, 3/b)$, *with probability at least* $1 - \delta$*:*

$$\forall n > 0 : \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \frac{\lambda \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i)}{2(1 - \lambda b/3)} + \frac{\ln(1/\delta)}{\lambda}.$$

*This implies that for all* $\sigma > 0$, *with probability at least* $1 - \delta$*:*

$$\forall n > 0 : \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \sigma \sqrt{2 \ln(1/\delta)} + \frac{b \ln(1/\delta)}{3}$$

$$\text{or } \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) > \sigma^2.$$

# Reference Used in the Proof of Theorem 8

## Lemma 9 (Lem 2.9 )

*Consider a random variable $X$ so that $\mathbb{E}[X] = \mu$. Assume that there exists $\alpha > 0$ and $\beta \geq 0$ such that for $\lambda \in [0, \beta^{-1})$:*

$$\Lambda_X(\lambda) \leq \lambda\mu + \frac{\alpha\lambda^2}{2(1 - \beta\lambda)}, \tag{1}$$

*then for $\epsilon > 0$:*

$$- I_X(\mu + \epsilon) \leq -\frac{\epsilon^2}{2(\alpha + \beta\epsilon)},$$

$$- I_X\left(\mu + \epsilon + \frac{\beta\epsilon^2}{2\alpha}\right) \leq -\frac{\epsilon^2}{2\alpha}.$$

# Proof of Theorem 8

Using the logarithmic moment generating function (2.13), we obtain the first inequality directly from the first inequality of Theorem 2. Moreover, we can obtain from the second inequality of Theorem 2 with

$$A_n = \left\{ \mathcal{S}_n : \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) \leq \sigma^2 \right\},$$

$$z_n = \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i - \epsilon - \epsilon^2 b/(6\sigma^2),$$

and the rate function estimate corresponding to the third inequality of Lemma 9:

$$\Pr\left[ \exists n > 0 : \sum_{i=1}^{n} \xi_i \leq z_n \text{ and } \mathcal{S}_n \in A_n \right] \leq \exp\left( -\frac{\epsilon^2}{2\sigma^2} \right).$$

This implies the second desired inequality with $\epsilon = \sigma\sqrt{2\ln(1/\delta)}$.

# Data Dependent Freedman's Inequality

We can remove the dependency on $\sigma$ in Theorem 8, which leads to the following result. One may also use the same technique to alleviate the dependence on $b$.

## Proposition 10

*Under the assumptions of Theorem 8, for $V_0 > 0$ and $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:*

$$\forall n > 0 : \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \sqrt{4 \left( V_0 + \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) \right) \ln((\hat{\ell} + 1)^2/\delta)} + \frac{b \ln((\hat{\ell} + 1)^2/\delta)}{3},$$

*where* $\hat{\ell} = \left\lfloor 1 + \log_2(1 + \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i)/V_0) \right\rfloor$.

# Proof of Proposition 10

We consider a sequence $\sigma^2 = 2^\ell V_0$ for $\ell = 1, 2, \ldots$. With probability at least $1 - \delta$, for all $\ell \geq 1$:

$$\forall n > 0 : \sum_{i=1}^{n} \mathbb{E}_{Z_i^{(y)}} \xi_i < \sum_{i=1}^{n} \xi_i + \sqrt{2^{\ell+1} V_0 \ln((\ell+1)^2/\delta)} + \frac{b \ln((\ell+11)^2/\delta)}{3}$$

$$\text{or} \quad \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) > 2^\ell V_0.$$

With $\hat{\ell} = \lfloor 1 + \log_2(1 + \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i)/V_0) \rfloor$, we have

$$\sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) \leq 2^{\hat{\ell}} V_0, \qquad 2^{\hat{\ell}+1} V_0 \leq 4 \left( V_0 + \sum_{i=1}^{n} \mathrm{Var}_{Z_i^{(y)}}(\xi_i) \right).$$

This implies the desired result.

# Uniform Convergence

Consider a real-valued function class $\mathcal{F}$ on $\mathcal{Z}$, and a sequence of observations $Z_1, \ldots, Z_n \in \mathcal{Z}$, and let

$$\mathcal{S}_n = [Z_1, \ldots, Z_n].$$

We assume that each $Z_t$ may depend on $\mathcal{S}_{t-1}$.

In uniform convergence, we are generally interested in estimating the following quantity

$$\sup_{f \in \mathcal{F}} \left[ \frac{1}{n} \sum_{i=1}^{n} [-f(Z_i) + \mathbb{E}_{Z_i^{(y)}} f(Z_i)] \right].$$

# Uniform Convergence with $L_\infty$ Packing Number

In the following theorem, $M(\epsilon, \mathcal{F}, \|\cdot\|_\infty))$ is the $\epsilon$ $L_\infty$ packing number of $\mathcal{F}$ with the metric $\|f\|_\infty = \sup_Z |f(Z)|$.

## Theorem 11 (Simplification of Thm 13.11)

*We have for any $\lambda > 0$, with probability at least $1 - \delta$, for all $n \geq 1$ and $f \in \mathcal{F}$:*

$$-\sum_{i=1}^{n} f(Z_i) - \frac{1}{\lambda} \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda f(Z_i)}$$
$$\leq \inf_{\epsilon > 0} \left[ 2n\epsilon + \frac{\ln(M(\epsilon, \mathcal{F}, \|\cdot\|_\infty))/\delta)}{\lambda} \right].$$

It is also possible to improve Theorem 11 using chaining (see Proposition 13.14).

# Proof of Theorem 11

Let $\mathcal{F}_\epsilon \subset \mathcal{F}$ be an $\epsilon$ maximal packing of $\mathcal{F}$, with $|\mathcal{F}_\epsilon| \leq M(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$. We obtain from Theorem 2, and the uniform bound over $\mathcal{F}_\epsilon$ that with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}_\epsilon} \left[ -\sum_{i=1}^{n} f(Z_i) - \frac{1}{\lambda} \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda f(Z_i)} \right] \leq \frac{\ln(M(\epsilon, \mathcal{F}, \|\cdot\|_\infty))/\delta)}{\lambda}.$$

Since $\mathcal{F}_\epsilon$ is also an $\epsilon$ $L_\infty$ cover of $\mathcal{F}$ (see Theorem 5.2), we obtain

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left[ -\sum_{i=1}^{n} f(Z_i) - \frac{1}{\lambda} \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda f(Z_i)} \right] \\ \leq &2n\epsilon + \sup_{f \in \mathcal{F}_\epsilon} \left[ -\sum_{i=1}^{n} f(Z_i) - \frac{1}{\lambda} \sum_{i=1}^{n} \ln \mathbb{E}_{Z_i^{(y)}} e^{-\lambda f(Z_i)} \right]. \end{aligned}$$

This implies the result.

# Refined Least Squares Uniform Convergence

## Theorem 12 (Simplification of Thm 13.15)

*Let $\{(X_t, \epsilon_t)\}$ be a filtered sequence in $\mathcal{X} \times \mathbb{R}$ so that $\epsilon_t$ is conditional zero-mean sub-Gaussian noise: for all $\lambda \in \mathbb{R}$,*

$$\ln \mathbb{E}[e^{\lambda \epsilon_t} | X_t, \mathcal{S}_{t-1}] \leq \frac{\lambda^2}{2} \sigma^2,$$

*where $\mathcal{S}_{t-1}$ denotes the history data. Assume that $Y_t = f_*(X_t) + \epsilon_t$, with $f_*(x) \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$. Let $\hat{f}_t$ be the exact ERM solution:*

$$\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t} (f(X_s) - Y_s)^2.$$

*Then with probability at least $1 - \delta$, for all $t \geq 0$:*

$$\sum_{s=1}^{t} (\hat{f}_t(X_t) - f_*(X_t))^2 \leq \inf_{\epsilon > 0} \left[ 8t\epsilon(\sigma + 2\epsilon) + 12\sigma^2 \ln \frac{2N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)}{\delta} \right].$$

# Minimax Analysis for Sequential Estimation

## Notations

Consider a general sequential estimation problem, where we observe data $Z_t \in \mathcal{Z}$ from the environment by interacting with the environment using a sequence of learned policies $\pi_t \in \Pi$.

At each time $t$, the observation history is

$$\mathcal{S}_{t-1} = [(Z_1, \pi_1), \ldots, (Z_{t-1}, \pi_{t-1})].$$

Based on the history, the player (or learning algorithm), denoted by $\hat{q}$, determines the next *policy* $\pi_t \in \Pi$ that can interact with the environment.

Based on the policy $\pi_t$, environment generates the next observation $Z_t \in \mathcal{Z}$ according to an unknown distribution $q(Z_t | \pi_t, \mathcal{S}_{t-1})$.

## Definition 13 (Sequential Statistical Estimation)

Consider a family of environment distributions $\mathcal{P}_{\mathcal{Z}}$, where each $q \in \mathcal{P}_{\mathcal{Z}}$ determines the probability for generating $Z_t$ based on policy $\pi_t$ as $q(Z_t|\pi_t, \mathcal{S}_{t-1})$. Consider also a family of learning algorithms, represented by $\mathcal{P}_{\mathcal{A}}$. Each learning algorithm $\hat{q} \in \mathcal{P}_{\mathcal{A}}$ maps the history $\mathcal{S}_{t-1}$ deterministically to the next policy $\pi_t \in \Pi$ as $\pi_t = \hat{q}(\mathcal{S}_{t-1})$. Given $q \in \mathcal{P}_{\mathcal{Z}}$, and $\hat{q} \in \mathcal{P}_{\mathcal{A}}$, the data generation probability is fully determined as

$$p(\mathcal{S}_n|\hat{q}, q) = \prod_{t=1}^{n} q(Z_t|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}).$$

After observing $\mathcal{S}_n$ for some $n$, the learning algorithm $\hat{q}$ determines a distribution $\hat{q}(\theta|\mathcal{S}_n)$, and draw estimator $\theta \in \Theta$ according to $\hat{q}(\theta|\mathcal{S}_n)$. The learning algorithm suffers a loss (also referred to as regret) $Q(\theta, q)$. The overall probability of $\theta$ and $\mathcal{S}_n$ is

$$p(\theta, \mathcal{S}_n|\hat{q}, q) = \hat{q}(\theta|\mathcal{S}_n) \prod_{t=1}^{n} q(Z_t|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}). \tag{2}$$

# Example: Online Learning (non-adversarial)

We consider a parameter space $\Omega$, and at each time $t$, the learning algorithm chooses a parameter $w_t \in \Omega$ according to a probability distribution $\pi_t(\cdot)$ on $\Omega$.

This probability distribution is the policy. Given $w_t \sim \pi_t$, we then observe a $Z_t \sim q_t$ from an unknown distribution $q_t$. We assume that the loss function $\ell$ is known.

After $n$ rounds, let $\theta(\mathcal{S}_n) = [\pi_1, \ldots, \pi_n]$, we suffer a loss

$$Q(\theta, q) = \sum_{t=1}^{n} \mathbb{E}_{w_t \sim \pi_t} \mathbb{E}_{Z_t \sim q_t} \ell(w_t, Z_t) - \inf_{w \in \Omega} \sum_{t=1}^{n} \mathbb{E}_{Z \sim q_t} \ell(w, Z_t).$$

# Example: MAB

We consider the multi-armed bandit problem, where we have $K$ arms from $\mathcal{A} = \{1, \dots, K\}$. For each arm $a \in \mathcal{A}$, we have a probability distribution $q_a$ on $[0, 1]$. If we pull an arm $a \in \mathcal{A}$, we observe a random reward $r \in [0, 1]$ from a distribution $q_a$ that depends on the arm $a$.

Our goal is to find the best arm $\theta \in \Theta = \mathcal{A}$ with the largest expected reward $\mathbb{E}_{r \sim q_a}[r]$, and the loss $Q(\theta, q) = \sup_a \mathbb{E}_{r \sim q_a}[r] - \mathbb{E}_{r \sim q_\theta}[r]$. In this case, a policy $\pi_t$ is a probability distribution over $\mathcal{A}$.

The learning algorithm defines a probability distribution $\hat{q}(\mathcal{S}_{t-1})$ over $\mathcal{A}$ at each time, and draw $a_t \sim \hat{q}(\mathcal{S}_{t-1})$. The observation $Z_t$ is the reward $r_t$ which is drawn from $q_{a_t}$.

# Contextual Bandits

In contextual bandits, we consider a context space $\mathcal{X}$ and action space $\mathcal{A}$. Given a context $x \in \mathcal{X}$, we can take an action $a \in \mathcal{A}$, and observe a reward $r \sim q_{x,a}$.

A policy $\pi$ is a map $\mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the set of probability distributions over $\mathcal{A}$ (with an appropriately defined sigma algebra).

The policy $\pi_t$ interacts with the environment to generate the next observation as: the environment generates $x_t$, the player takes an action $a_t \sim \pi_t(x_t)$, and then the environment generates the reward $r_t \sim q_{x_t,a_t}$.

# Minimax Risk

**Definition 14**

Consider an environment distribution family $\mathcal{P}_{\mathcal{Z}}$, learning algorithm distribution family $\mathcal{P}_{\mathcal{A}}$. Then the worst case expected risk of a learning algorithm $\hat{q} \in \mathcal{P}_{\mathcal{A}}$ with respect to $\mathcal{P}_{\mathcal{Z}}$ is given by

$$r_n(\hat{q}, \mathcal{P}_{\mathcal{Z}}, Q) = \sup_{q \in \mathcal{P}_{\mathcal{Z}}} \mathbb{E}_{\theta, \mathcal{S}_n \sim p(\cdot|\hat{q}, q)} Q(\theta, q),$$

where $p(\cdot|\hat{q}, q)$ is defined in (2). Moreover, the minimax risk is defined as:

$$r_n(\mathcal{P}_{\mathcal{A}}, \mathcal{P}_{\mathcal{Z}}, Q) = \inf_{\hat{q} \in \mathcal{P}_{\mathcal{A}}} r_n(\hat{q}, \mathcal{P}_{\mathcal{Z}}, Q).$$

# Lower Bound based on Assouad's Lemma

## Theorem 15 (Thm 13.24)

*Let $d \geq 1$ and $m \geq 2$ be integers, and let $\mathcal{P}_{\mathcal{Z}} = \{q^\tau : \tau \in \{1, \ldots, m\}^d\}$ contain $m^d$ probability measures. Suppose that the loss function $Q$ can be decomposed as $Q(\theta, q) = \sum_{j=1}^d Q_j(\theta, q)$, where $Q_j \geq 0$ are all non-negative. For each $j$, $\tau \sim_j \tau'$ if $\tau = \tau'$ or if $\tau$ and $\tau'$ differs by only one component $j$. Assume that there exists $\epsilon, \beta \geq 0$ such that*

$$\forall \tau' \sim_j \tau, \tau' \neq \tau : \quad [Q_j(\theta, q^\tau) + Q_j(\theta, q^{\tau'})] \geq \epsilon,$$

*and there exists $q_j^\tau$ such that all $\tau' \sim_j \tau$ map to the same value: $q_j^{\tau'} = q_j^\tau$. Given any learning algorithm $\hat{q}$. If for all $\tau$, $j \in [d]$, time step $t$, and $\mathcal{S}_{t-1}$:*

$$\frac{1}{m} \sum_{\tau' \sim_j \tau} \mathrm{KL}(q_j^\tau(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}) || q^{\tau'}(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}))) \leq \beta_{j,t}^2, \qquad \textit{then}$$

$$\frac{1}{m^d} \sum_{\tau} \mathbb{E}_{\theta, \mathcal{S}_n \sim p(\cdot|\hat{q}, q^\tau)} Q(\theta, q^\tau) \geq 0.5 d\epsilon \left(1 - \sqrt{\frac{2}{d} \sum_{j=1}^{d} \sum_{t=1}^{n} \beta_{j,t}^2}\right).$$

# Result used in the Proof of Theorem 15

## Lemma 16 (Generalized Assouad's Lemma, Lem 12.27 )

*Consider a finite family of distributions $\mathcal{P}$. Let $d \geq 1$ be an integer, and $Q$ can be decomposed as*

$$Q(\theta, \mathcal{D}) = \sum_{j=1}^{d} Q_j(\theta, \mathcal{D}),$$

*where $Q_j \geq 0$ are all non-negative. Assume for all $j$, there exists a partition $M_j$ of $\mathcal{P}$. We use notation $\mathcal{D}' \sim_j \mathcal{D}$ to indicate that $\mathcal{D}'$ and $\mathcal{D}$ belong to the same partition in $M_j$. Let $m_j(\mathcal{D})$ be the number of elements in the partition containing $\mathcal{D}$. Assume there exist $\epsilon, \beta \geq 0$ such that*

$$\forall \mathcal{D}' \sim_j \mathcal{D}, \mathcal{D}' \neq \mathcal{D} : \quad \inf_{\theta} [Q_j(\theta, \mathcal{D}') + Q_j(\theta, \mathcal{D})] \geq \epsilon,$$

$$\forall \mathcal{D} \in \mathcal{P} : \quad \frac{1}{d(\mathcal{P})} \sum_{j=1}^{d} \sum_{\mathcal{D} \in \mathcal{P}_j} \frac{1}{m_j(\mathcal{D}) - 1} \sum_{\mathcal{D}' \sim_j \mathcal{D}} \|\mathcal{D}' - \mathcal{D}\|_{\mathrm{TV}} \leq \beta,$$

*where $\mathcal{P}_j = \{\mathcal{D} \in \mathcal{P} : m_j(\mathcal{D}) > 1\}$ and $d(\mathcal{P}) = \sum_{j=1}^{d} |\mathcal{P}_j|$. Let $\mathcal{A}(Z)$ be any estimator, we have*

$$\frac{1}{|\mathcal{P}|} \sum_{\mathcal{D} \in \mathcal{P}} \mathbb{E}_{Z \sim \mathcal{D}} \mathbb{E}_{\mathcal{A}} Q(\mathcal{A}(Z), \mathcal{D}) \geq \frac{\epsilon d(\mathcal{P})}{2|\mathcal{P}|} \left[1 - \beta\right],$$

*where $\mathbb{E}_{\mathcal{A}}$ is with respect to the internal randomization in $\mathcal{A}$.*

# Proof of Theorem 15 (I/II)

We have

$$
\begin{aligned}
&\frac{1}{dm^d}\sum_{\tau}\sum_{j=1}^{d}\frac{1}{m-1}\sum_{\tau'\sim_j\tau}\|p(\cdot|\hat{q},q^{\tau})-p(\cdot|\hat{q},q^{\tau'})\|_{\mathrm{TV}}\\
\leq&\frac{1}{dm^d}\sum_{\tau}\sum_{j=1}^{d}\frac{1}{m-1}\sum_{\tau'\sim_j\tau,\tau'\neq\tau}\Big[\|p(\cdot|\hat{q},q_j^{\tau})-p(\cdot|\hat{q},q^{\tau'})\|_{\mathrm{TV}}\\
&\qquad+\|p(\cdot|\hat{q},q_j^{\tau})-p(\cdot|\hat{q},q^{\tau})\|_{\mathrm{TV}}\Big]\\
=&\frac{2}{dm^d}\sum_{\tau}\sum_{j=1}^{d}\frac{1}{m}\sum_{\tau'\sim_j\tau}\|p(\cdot|\hat{q},q_j^{\tau})-p(\cdot|\hat{q},q^{\tau'})\|_{\mathrm{TV}}=A.
\end{aligned}
$$

The first inequality is triangle inequality for $\mathrm{TV}$-norm. The first equality used $\hat{q}_j^{\tau}=\hat{q}_j^{\tau'}$ when $\tau\sim_j\tau'$.

# Proof of Theorem 15 (II/II)

$$A \leq \frac{2}{m^d} \sum_{\tau} \sqrt{\frac{1}{dm} \sum_{j=1}^{d} \sum_{\tau' \sim_j \tau} \| p(\cdot|\hat{q}, q_j^{\tau}) - p(\cdot|\hat{q}, q^{\tau'}) \|_{\mathrm{TV}}^2}$$

(Jensen's inequality for $\sqrt{\cdot}$)

$$\leq \frac{2}{m^d} \sum_{\tau} \sqrt{\frac{1}{2dm} \sum_{j=1}^{d} \sum_{\tau' \sim_j \tau} \mathrm{KL}(p(\cdot|\hat{q}, q_j^{\tau}) || p(\cdot|\hat{q}, q^{\tau'}))}$$

(Pinsker's inequality)

$$\leq \sqrt{\frac{2}{d} \sum_{j=1}^{d} \sum_{t=1}^{n} \beta_t^2}.$$

The last inequality used Lemma 13.21. Now in Lemma 16, we let $M_j(q^{\tau}) = \{q^{\tau'} : \tau' \sim_j \tau\}$ be the partitions. The result is a simple application of Lemma 16 with $m_j(q^{\tau}) = m$, $|\mathcal{P}| = m^d$, and $d(\mathcal{P}) = dm^d$.

# Example

Consider estimating the mean of a $d$ dimensional Gaussian random variable $Z \sim N(\theta, I_{d\times d})$. Each time the player draws an action $a_t \in \{1, \dots, d\}$, and the environment draws $\tilde{Z}_t \sim N(\theta, I_{d\times d})$, and reveals only the $a_t$-th component $Z_t = \tilde{Z}_{t,a_t}$. After $T$ rounds, we would like to estimate the mean as $\hat{\theta}$, and measure the quality with $Q(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$. In this case, a policy $\pi_t$ can be regarded as a distribution over $\{1, \dots, d\}$, and we draw $a_t \sim \pi_t$.
To obtain an upper bound of the loss, we can simply randomly pick $a_t$, and use the following unbiased estimator:

$$\hat{\theta}_j = \frac{d}{n} \sum_{t=1}^{n} Z_{t,a_t} \mathbb{1}(a_t = j).$$

This implies that

$$\mathbb{E} \, \|\hat{\theta} - \theta\|_2^2 = \frac{d^2}{n}.$$

# Example (cont)

To obtain a lower bound of the loss, we consider Corollary 13.23, with $\theta^\tau = \epsilon\tau/(\sqrt{d})$ and $\mathcal{P}_\mathcal{Z} = \{N(\theta^\tau, I_{d\times d}) : \tau \in \{\pm 1\}^d$. Consider the decomposition

$$Q(\theta, q^\tau) = \sum_{j=1}^{d} Q_j(\theta, q^\tau), \quad Q_j(\theta, q^\tau) = (\theta_j - \theta_j^\tau)^2.$$

This implies that

$$\forall \tau : \quad [Q_j(\theta, q^\tau) + Q_j(\theta, q^{\tau^{-[j]}})] \geq \epsilon^2/d.$$

Let $Z_t$ and $Z_t'$ be the observations under $q, q' \in \mathcal{P}_\mathcal{Z}$, then for any $a_t$, $\mathrm{KL}(Z_t, Z_t') \leq \beta_t^2 = 2\epsilon^2/d$. When

$$2n\epsilon^2 \leq d^2/32 - d,$$

we have

$$r_n(\mathcal{P}_\mathcal{A}, \mathcal{P}_\mathcal{Z}, Q) \geq \epsilon^2/16.$$

This matches the upper bound up to a constant.

# Summary (Chapter 13)

- Martingale Exponential Equality
- Martingale Exponential Tail Probability Inequality
- Azuma's Inequality
- Freedman's Inequality
- Data Dependent Bound
- Uniform Convergence with $L_\infty$ Packing Number
- Minimax Analysis and Lower Bound