Lecture 15

ERM, Uniform Convergence

Sasha Rakhlin

Oct 29, 2018

Outline

Outline

Recall that we proved

$$\mathbb{E}\mathbf{L}(w_T) \leq \frac{1}{n+1} \times \frac{D^2}{\gamma^2}$$

for the last hyperplane w_T of Perceptron (cycled until no more mistakes) under the assumption that there exists w^* with $\|w^*\| = 1$ such that $Y\langle w^*, X\rangle \geq \gamma$ and $D \geq \|X\|$ almost surely.

This is a result about a particular minimizer of empirical loss $\widehat{\mathbf{L}}_{01}(w)$.

Full gradient for i.i.d. data

Suppose that instead of multi-pass Perceptron, we run full gradient descent on empirical objective

$$\widehat{\mathbf{L}}(w) = \frac{1}{n} \sum_{t=1}^{n} \max \{-Y_t \langle w, X_t \rangle, 0\}$$

NB: hinge-at-zero $\max\{-y \langle w, x \rangle, 0\}$ is not a surrogate loss for the indicator loss, but its minimizer does enjoy zero indicator loss.

If all we know is that w minimizes empirical loss $\widehat{\mathbf{L}}(w)$ (but not necessarily obtained via Perceptron), what can we do?

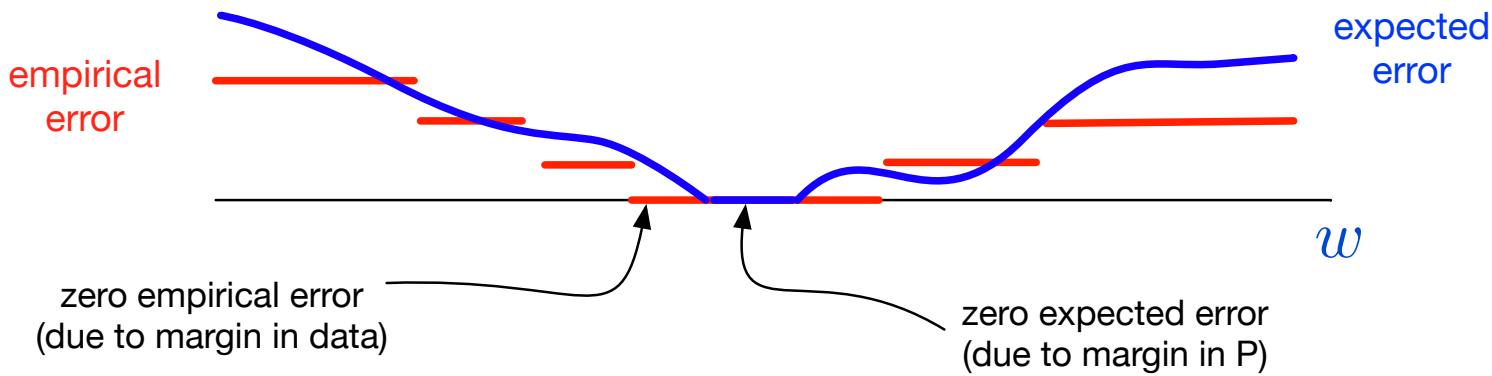
Beyond hyperplanes with margin, what can we say about expected loss of an empirical minimizer?

It is useful to consider the function

$$w \mapsto \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{Y_t \langle w, X_t \rangle \leq 0\}$$

as w varies (but fixing the data). For example, values of this function over the sphere $\|w\| = 1$ look like [draw in class], while the expected error looks like [draw in class].

If we unroll this picture of $\mathbf{L}(w)$ and $\widehat{\mathbf{L}}(w)$, it would look like this:



Blue curve is fixed (given P) while red curve changes according to a draw of data.

Interpret our claim about last-step-Perceptron as: there is a choice of empirical minimizer ($\widehat{\mathbf{L}}(w_{\mathrm{T}}) = 0$) such that in expectation (over draw of data) its out-of-sample performance (blue curve) is $\mathrm{O}(1/\mathrm{n})$.

Do we expect any minimizer of the empirical profile to have a small expected error?

Let's make things more general. Fix some loss function ℓ and a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$. *Empirical Risk Minimization* (ERM) algorithm is

$$\widehat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathbf{L}}(f)$$

In the linear case,

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : \|w\| \leq 1\}$$

Performance of ERM

If $\widehat{f}_n$ is an ERM,

$$\begin{aligned}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) &= \left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\} + \left\{\widehat{\mathbf{L}}(\widehat{f}_n) - \widehat{\mathbf{L}}(f_{\mathcal{F}})\right\} + \left\{\widehat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\} \\ &\leq \left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\} + \left\{\widehat{\mathbf{L}}(f_{\mathcal{F}}) - \mathbf{L}(f_{\mathcal{F}})\right\}\end{aligned}$$

because the second term is negative.

Now take expectation on both sides and observe that second term is zero in expectation:

$$\mathbb{E}_{\mathcal{S}}\left[\widehat{\mathbf{L}}(f_{\mathcal{F}})\right] - \mathbf{L}(f_{\mathcal{F}}) = 0$$

So, estimation error

$$\mathbb{E}_{\mathcal{S}}\left[\mathbf{L}(\widehat{f}_n)\right] - \mathbf{L}(f_{\mathcal{F}}) \leq \mathbb{E}_{\mathcal{S}}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right]$$

Let's do that last step slowly: for any fixed (data-independent) function f (including $f_{\mathcal{F}}$)

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left\{\widehat{\mathbf{L}}(f)\right\} &= \mathbb{E}_{\mathcal{S}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}}\left\{\ell(f(X_i), Y_i)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(X,Y)}\left\{\ell(f(X), Y)\right\} \\
&= \mathbf{L}(f)
\end{aligned}$$

Wait, Are We Done?

Can't we also apply above calculation to show that

$$\mathbb{E}_{\mathcal{S}}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right]$$

is zero? No, because $\widehat{f}_n$ is data-dependent:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left\{\widehat{\mathbf{L}}(\widehat{f}_n)\right\} &= \mathbb{E}_{\mathcal{S}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(\widehat{f}_n(X_i), Y_i)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}}\left\{\ell(\widehat{f}_n(X_i), Y_i)\right\} \\
&= \mathbb{E}_{\mathcal{S}}\left\{\ell(\widehat{f}_n(X_i), Y_i)\right\} \\
&\neq \mathbb{E}_{\mathcal{S},(X,Y)}\left\{\ell(\widehat{f}_n(X), Y)\right\}
\end{aligned}$$

The next-to-last term is "in sample" while expected loss is "out of sample."

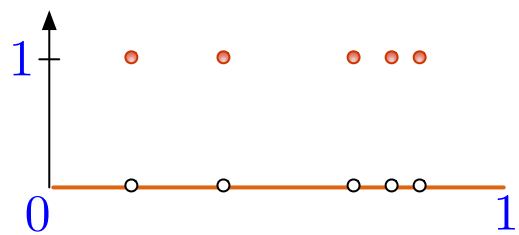
We say that $\ell(\widehat{f}_n(X_i), Y_i)$ is a *biased estimate* of $\mathbb{E}\ell(\widehat{f}_n(X), Y)$.

How bad can this bias be?

Example

- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$
- $\ell(f(X_i), Y_i) = \mathbf{I}\{f(X_i) \neq Y_i\}$
- distribution $P = P_X \times P_{Y|X}$ with $P_X = \text{Unif}[0,1]$ and $P_{Y|X} = \delta_{Y=1}$
- function class

$$\mathcal{F} = \cup_{n \in \mathbb{N}} \left\{ f = f_S : S \subset \mathcal{X}, |S| = n, f_S(x) = \mathbf{I}\{x \in S\} \right\}$$



ERM $\widehat{f}_n$ memorizes (perfectly fits) the data, but has no ability to generalize. Observe that

$$0 = \mathbb{E}\ell(\widehat{f}_n(X_i), Y_i) \quad \neq \quad \mathbb{E}\ell(\widehat{f}_n(X), Y) = 1$$

This phenomenon is called *overfitting* (though the term is vague and used in a variety of ways).

Example

NB: we will see later in the course that memorization methods can generalize, so in the previous example it was not just the memorization part that was problematic, but the overall definition of $\widehat{f}_n$.

In fact, we already saw that perfectly fitting the data (at least in terms of perfectly separating the dataset) did not prevent Perceptron's last output to have good generalization.

Example

Where do we go from here? Two approaches:

1. uniform deviations (remove the hat altogether)
2. find properties of algorithms that limit in some way the bias of $\ell(\widehat{f}_n(X_i), Y_i)$. *Stability*, *differential privacy*, *compression* are such approaches.

Outline

Removing the Hat

Recall: the difficulty in bounding the generalization gap was that $\widehat{f}_n$ depends on the data. The key trick here is to remove the dependence by "maxing out":

$$\mathbb{E}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right] \leq \mathbb{E} \max_{f \in \mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right]$$

For this inequality to hold, we only need to know that $\widehat{f}_n$ takes values in $\mathcal{F}$ (does not have to be ERM).

If we have some extra knowledge on the location of $\widehat{f}_n$, we should take supremum over that (ideally, data-independent) subset of $\mathcal{F}$. This is called "localized analysis."

Uniform Deviations

We first focus on understanding

$$\mathbb{E} \max_{f \in \mathcal{F}} \left\{ \mathbb{E}_{X,Y} \ell(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right\}.$$

If $\mathcal{F} = \{f_0\}$ consists of a single function, then above is 0. However, if $\mathcal{F} = \{f_0, f_1\}$, above is $O(1/\sqrt{n})$ as soon as f_0 and f_1 "different enough."

A bit of notation to simplify things...

To ease the notation,

- Let $z_i = (x_i, y_i)$ so that the training data is $\{z_1, \ldots, z_n\}$
- $g(z) = \ell(f(x), y)$ for $z = (x, y)$
- Loss class $\mathcal{G} = \{g : g(z) = \ell(f(x), y)\} = \ell \circ \mathcal{F}$
- $\hat{g}_n = \ell(\widehat{f}_n(\cdot), \cdot)$, $g_{\mathcal{G}} = \ell(f_{\mathcal{F}}(\cdot), \cdot)$
- $g^* = \arg\min_g \mathbb{E}g(z) = \ell(f^*(\cdot), \cdot)$ is Bayes optimal (loss) function

We can now work with the set $\mathcal{G}$, but keep in mind that each $g \in \mathcal{G}$ corresponds to an $f \in \mathcal{F}$:

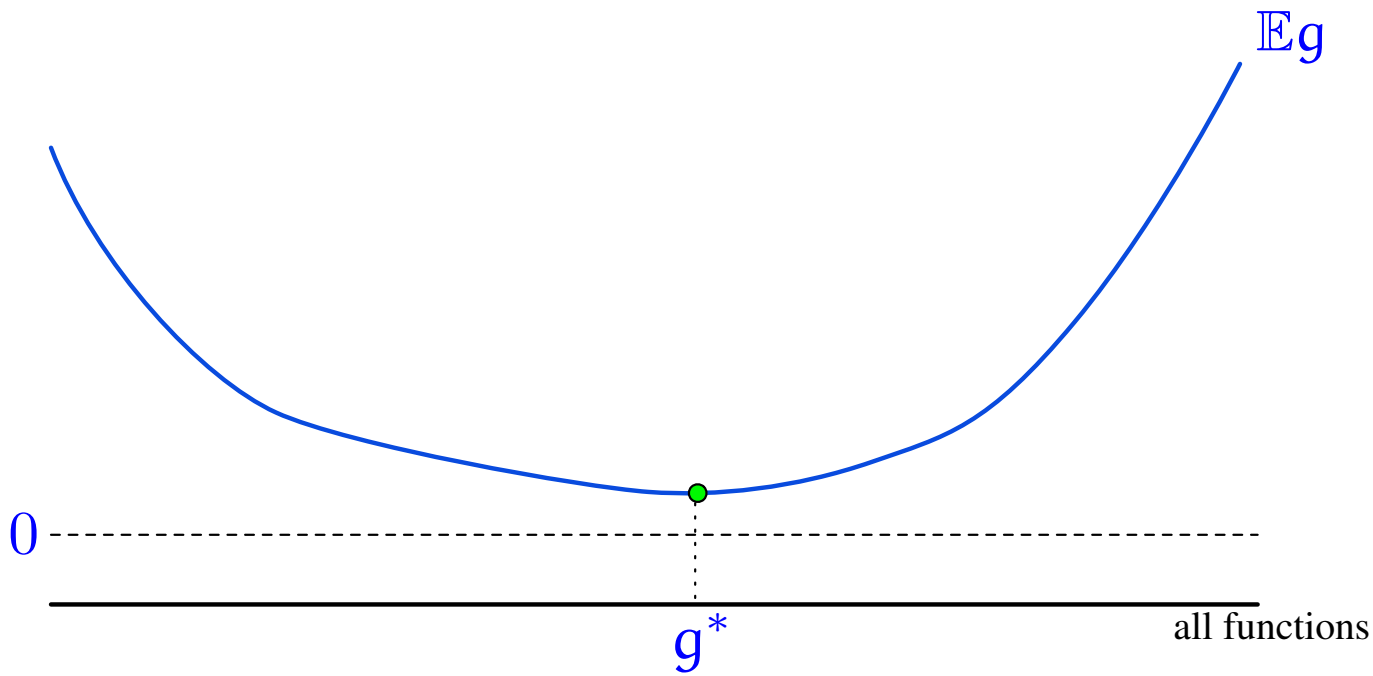
$$g \in \mathcal{G} \quad \longleftrightarrow \quad f \in \mathcal{F}$$

Once again, the quantity of interest is

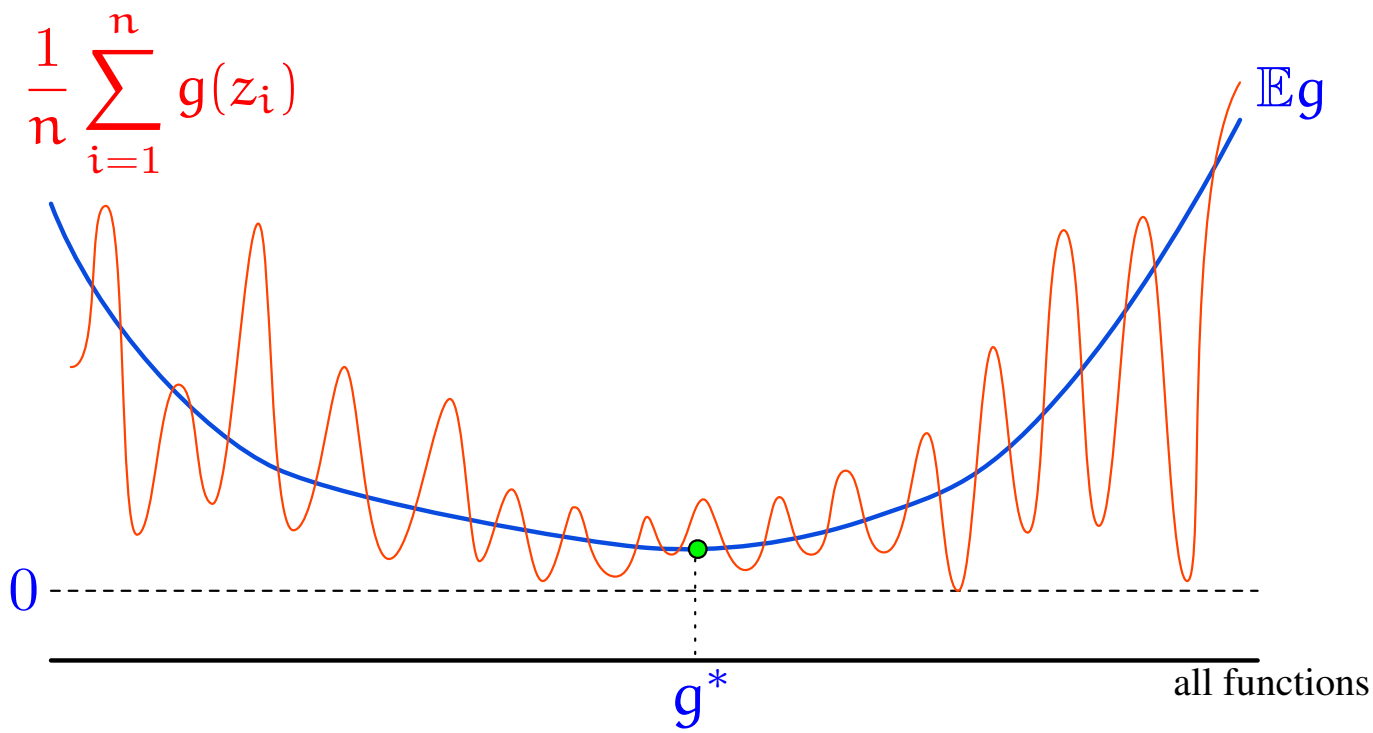
$$\max_{g \in \mathcal{G}} \left\{ \mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\}$$

On the next slide, we visualize deviations $\mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)$ for all possible functions g and discuss all the concepts introduces so far.

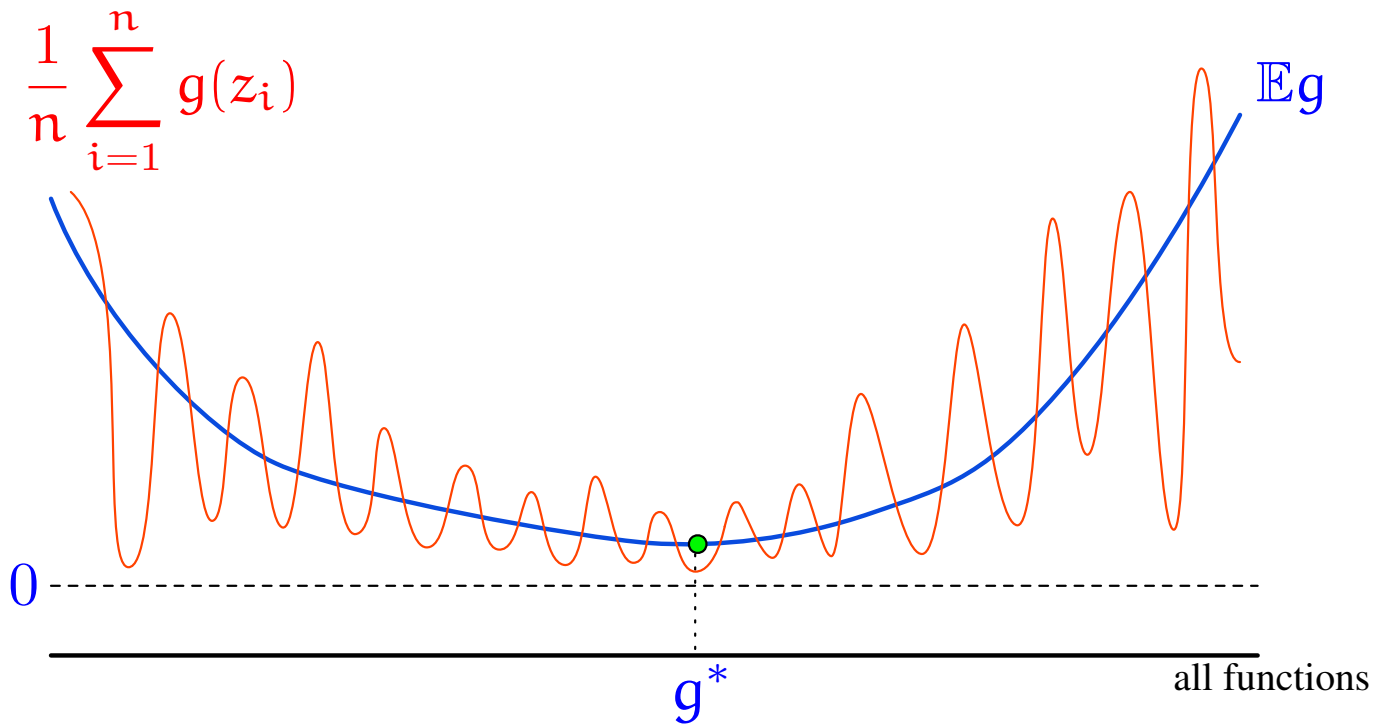
Empirical Process Viewpoint



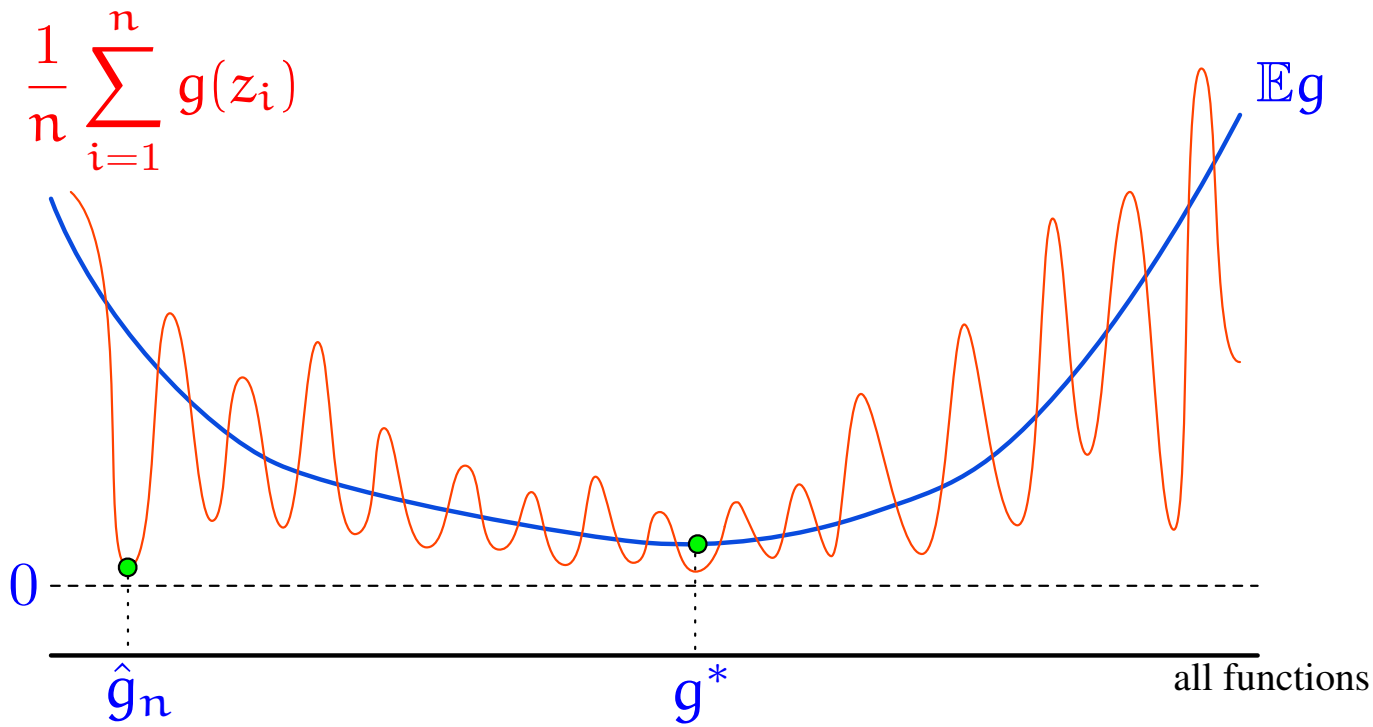
Empirical Process Viewpoint



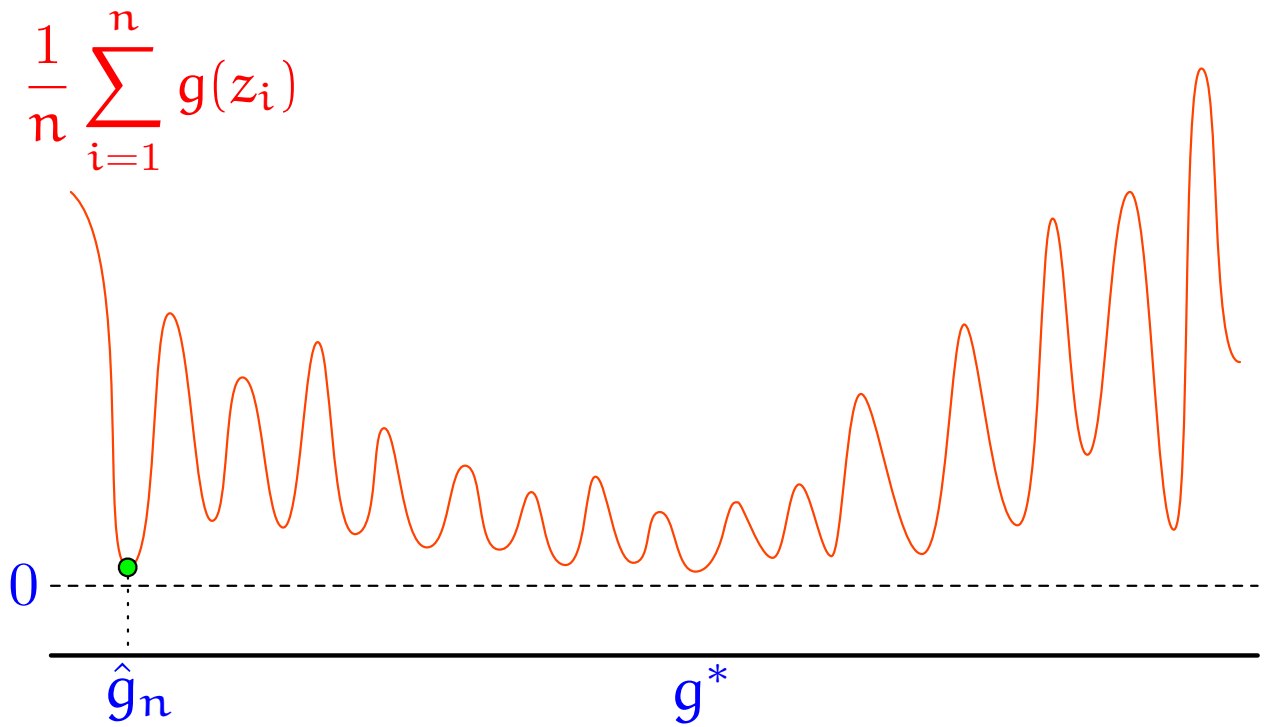
Empirical Process Viewpoint



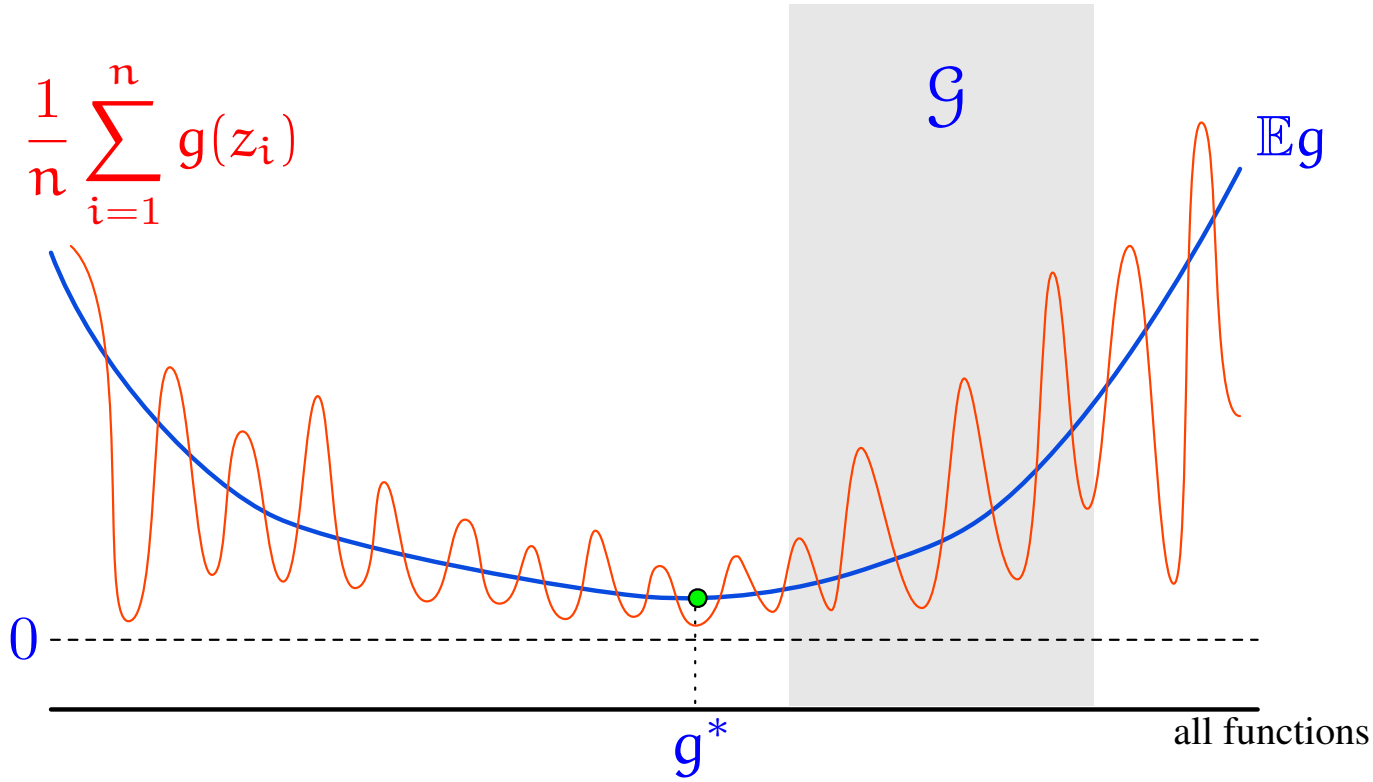
Empirical Process Viewpoint



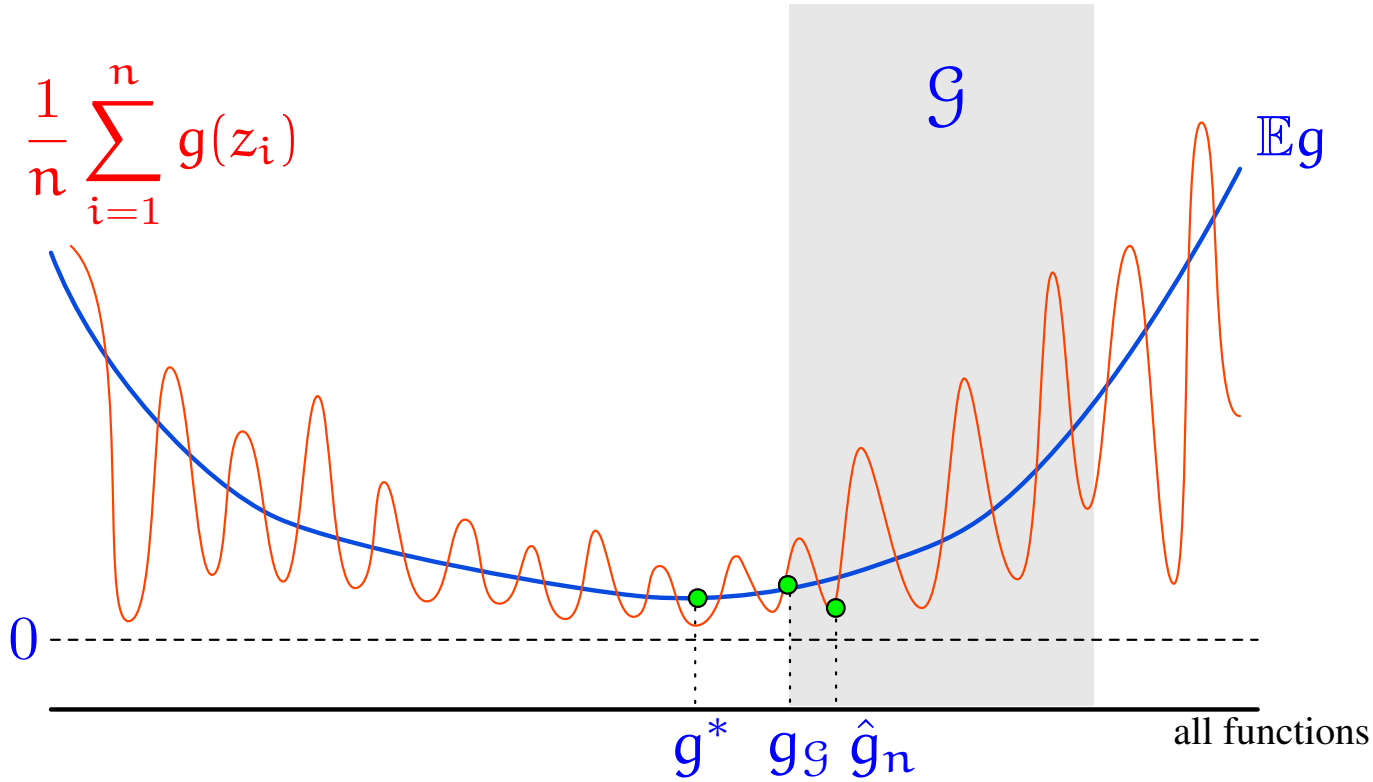
Empirical Process Viewpoint



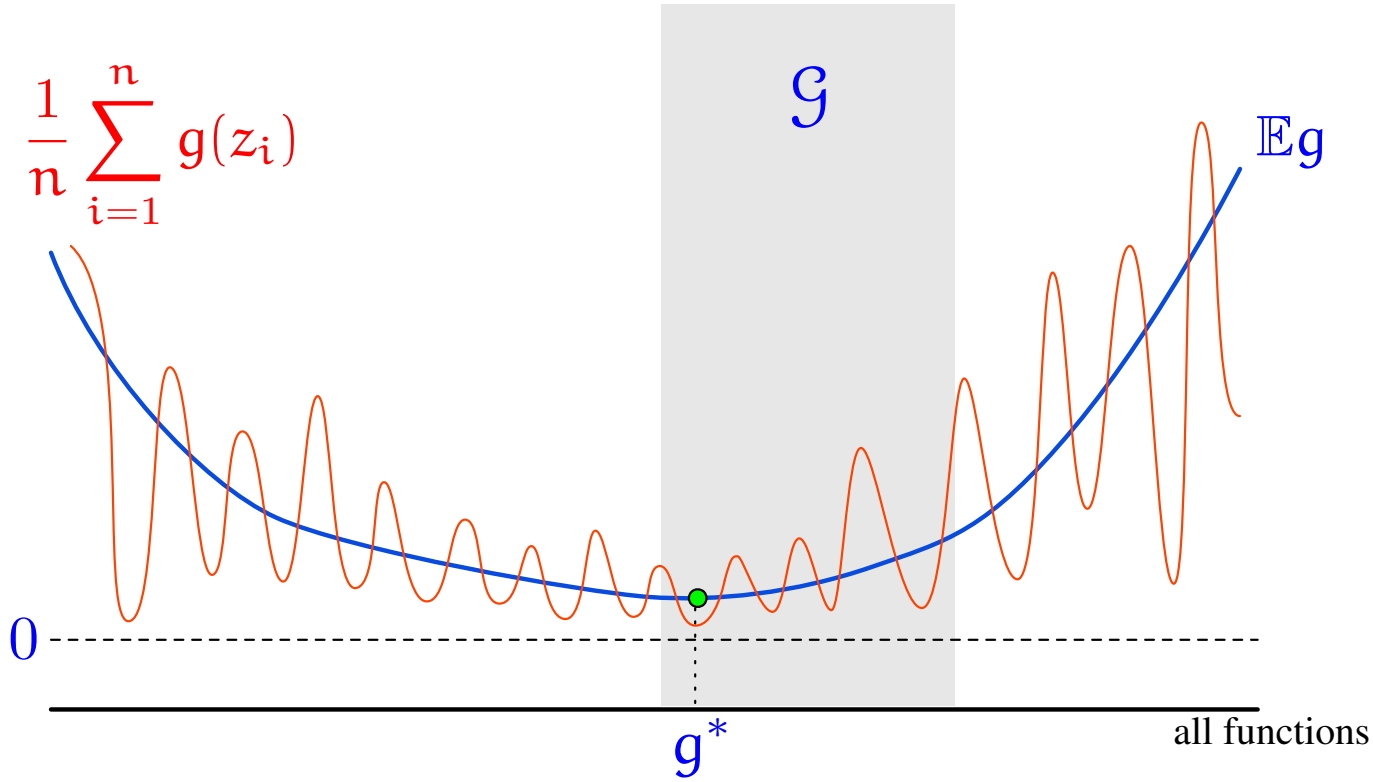
Empirical Process Viewpoint



Empirical Process Viewpoint



Empirical Process Viewpoint



Empirical Process Viewpoint

A *stochastic process* is a collection of random variables indexed by some set.

An *empirical process* is a stochastic process

$$\left\{ \mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\}_{g \in \mathcal{G}}$$

indexed by a function class $\mathcal{G}$.

Uniform Law of Large Numbers:

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}g - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \to 0$$

in probability.

Empirical Process Viewpoint

A *stochastic process* is a collection of random variables indexed by some set.

An *empirical process* is a stochastic process

$$\left\{ \mathbb{E}g(z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right\}_{g \in \mathcal{G}}$$

indexed by a function class $\mathcal{G}$.

Uniform Law of Large Numbers:

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}g - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \to 0$$

in probability.

Key question: How "big" can $\mathcal{G}$ be for the supremum of the empirical process to still be manageable?

Important distinction:

A take-away message is that the following two statements are worlds apart:

with probability at least $1-\delta$, for any $g \in \mathcal{G}$, $\quad \mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \le \epsilon$

vs

for any $g \in \mathcal{G}$, with probability at least $1-\delta$, $\quad \mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \le \epsilon$

The second statement follows from CLT, while the first statement is often difficult to obtain and only holds for some $\mathcal{G}$.