Lecture 10: Posterior Convergence in Nonparametric Regression

Thibault Randrianarisoa

UTSC

April 2, 2026

The nonparametric regression model

We observe n data points from the nonparametric regression model:

$$Y_i = f(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $f : [0,1] \to \mathbb{R}$ is the unknown regression function, $t_1, \ldots, t_n \in [0,1]$ are fixed design points, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

We measure performance using the empirical L^2-norm:

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f(t_i)^2, \qquad d_n(f,g) = \|f - g\|_n.$$

The Bayesian approach places a prior Π on f and sets:

$$f \sim \Pi, \qquad Y^{(n)} = (Y_1, \ldots, Y_n) \,|\, f \sim P_f^{(n)}.$$

Bayes' formula for the posterior

Notation. Write $Y = Y^{(n)} = (Y_1, \ldots, Y_n)$, $p_f = p_f^{(n)}$. Denote by E_f the expectation under P_f.

Bayes' formula gives the posterior mass of any measurable set B:

$$\Pi[B \mid Y] = \frac{\int_B p_f(Y)\, d\Pi(f)}{\int p_f(Y)\, d\Pi(f)}.$$

Key observation: If $\Pi[B] = 0$, then $\Pi[B \mid Y] = 0$.

Frequentist analysis: Assume a true regression function f_0 with $Y_i = f_0(t_i) + \varepsilon_i$, and study $\Pi[\cdot \mid Y]$ in probability under P_{f_0}.

Dividing by $p_{f_0}(Y)$, rewrite as:

$$\Pi[B \mid Y] = \int_B \frac{p_f}{p_{f_0}}(Y)\, d\Pi(f) \Big/ \int \frac{p_f}{p_{f_0}}(Y)\, d\Pi(f).$$

Contraction rate

A sequence ε_n (with $\varepsilon_n \to 0$ as $n \to \infty$) is a contraction rate around f_0 for $\Pi[\,\cdot \mid Y]$ if, as $n \to \infty$,

$$E_{f_0} \Pi\big[\|f - f_0\|_n > \varepsilon_n \mid Y\big] = o(1).$$

We then wonder about events $B = \{f \,:\, \|f - f_0\|_n > \varepsilon_n\}$.

What target rate? Depends on the smoothness of f_0. For $\alpha > 0$, define the Hölder ball

$$C^\alpha(L) = \{g : [0,1] \to \mathbb{R} \,:\, \forall\, x, y \in [0,1],\ |g^{(\lfloor\alpha\rfloor)}(x) - g^{(\lfloor\alpha\rfloor)}(y)| \leqslant L|x-y|^{\alpha - \lfloor\alpha\rfloor}\},$$

where $\lfloor\alpha\rfloor$ is the integer part.

The minimax rate over $\mathcal{F}$ is of order $n^{-\alpha/(2\alpha+1)}$ (up to logarithmic factors):

$$\bar{\varepsilon}_n = \inf_T \sup_{f \in \mathcal{F}} E_f \, \|T - f\|_n \asymp n^{-\alpha/(2\alpha+1)}.$$

Kullback–Leibler divergence and its variance

For two probability measures P and Q with densities p, q, define the Kullback–Leibler (KL) divergence:

$$KL(P||Q) = \int p \log \frac{p}{q},$$

and its variance:

$$V(P||Q) = \int p \left(\log \frac{p}{q} - KL(P||Q) \right)^2.$$

KL-type neighborhood

For any $\varepsilon > 0$, define

$$B_K(f_0, \varepsilon) = \left\{ f \ : \ KL(P_{f_0}||P_f) \leqslant n\varepsilon^2, \ \ V(P_{f_0}||P_f) \leqslant n\varepsilon^2 \right\}.$$

KL neighborhood in the regression model

In the regression model $Y_i = f(t_i) + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, the log-likelihood ratio is

$$\log \frac{p_f}{p_g}(Y) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (f(t_i) - g(t_i)) \Big(Y_i - \frac{f(t_i) + g(t_i)}{2} \Big).$$

A direct computation gives:

$$KL(P_f || P_g) = \frac{n}{2\sigma^2} \|f - g\|_n^2, \qquad V(P_f || P_g) = \frac{n^2}{\sigma^4} \|f - g\|_n^4.$$

Taking $\sigma = 1$ for simplicity, the KL-neighborhood becomes:

$$B_K(f_0, \varepsilon) = \big\{ f : \|f - f_0\|_n^2 \leqslant 2\varepsilon^2 \ \text{ and } \ n\|f - f_0\|_n^4 \leqslant \varepsilon^2 \big\},$$

which, for $n\varepsilon^2 \to \infty$, is essentially the empirical L^2-ball $\{f : \|f - f_0\|_n \lesssim \varepsilon\}$.

Bounding the denominator

This neighborhood of f_0 plays a key role in bounding the denominator of Bayes' formula from below.

Lemma

For any prior Π on $L^2[0,1]$, for any $C, \varepsilon > 0$, with P_{f_0}-probability at least $1 - 1/(C^2 n\varepsilon^2)$,

$$\int \frac{p_f}{p_{f_0}}(Y)\, d\Pi(f) \;\geqslant\; \Pi[B_K(f_0, \varepsilon)]\, e^{-(1+C)n\varepsilon^2}.$$

Proof idea:

- Restrict integration to $B := B_K(f_0, \varepsilon)$ and use $\overline{\Pi}(\cdot) = \Pi(\cdot \cap B)/\Pi(B)$
- Apply Jensen's inequality to get a lower bound via the log-likelihood ratio
- Use Chebyshev's inequality with the KL variance to control the remainder
- Conclude by taking exponentials and renormalising

Proof (sketch)

Let $B := B_K(f_0, \varepsilon)$, suppose $\Pi(B) > 0$, and set $\overline{\Pi}(\cdot) = \Pi(\cdot \cap B)/\Pi(B)$. Then:

$$\int \frac{p_f}{p_{f_0}}(Y)\, d\Pi(f) \;\geqslant\; \Pi(B) \int \frac{p_f}{p_{f_0}}(Y)\, d\overline{\Pi}(f).$$

By Jensen's inequality:

$$\log \int \frac{p_f}{p_{f_0}}(Y)\, d\overline{\Pi}(f) \;\geqslant\; \int_B \log \frac{p_f}{p_{f_0}}(Y)\, d\overline{\Pi}(f) \;\geqslant\; -Z - n\varepsilon^2,$$

where $Z := \int_B \left[\log \frac{p_{f_0}}{p_f}(Y) - K(P_{f_0}, P_f) \right] d\overline{\Pi}(f)$.

Define $\mathcal{B}_n := \{|Z| \leqslant Cn\varepsilon^2\}$. By Chebyshev: $P_{f_0}[\mathcal{B}_n^c] \leqslant 1/(C^2 n\varepsilon^2)$.
On $\mathcal{B}_n$: $\log \int \frac{p_f}{p_{f_0}}(Y)\, d\overline{\Pi}(f) \geqslant -(C+1)n\varepsilon^2$, giving the result.

Lemma

Let $A_n \subset L^2[0,1]$ be measurable. If ε_n verifies $n\varepsilon_n^2 \to \infty$ and

$$\frac{\Pi[A_n]}{e^{-2n\varepsilon_n^2}\,\Pi[B_K(f_0,\varepsilon_n)]} = o(1),$$

then, as $n \to \infty$,

$$E_{f_0}\,\Pi[A_n \mid Y] = o(1).$$

Message: If the prior puts very little mass on a (sequence of) set(s) A_n relative to its mass near f_0, then the posterior also puts little mass on A_n.

Proof sketch: Write $\Pi[A_n|Y] = N/D$. Bound D from below using the previous Lemma. Bound N/D from above, take expectations, and use Fubini with $E_{f_0}[p_f/p_{f_0}(Y)] \leqslant 1$.

Why does this matter? Parametric vs. nonparametric

The Lemma requires checking that the ratio $\dfrac{\Pi[A_n]}{e^{-2n\varepsilon_n^2}\,\Pi[B_K(f_0,\varepsilon_n)]} \to 0$.

Parametric setting ($\theta \in \mathbb{R}^d$): prior masses of balls scale polynomially in ε_n. For instance, if Π has a continuous positive density near θ_0 and $B_K(f_0,\varepsilon_n)$ behaves like a ball:

$$\Pi[A_n] \asymp \varepsilon_n^d, \qquad \Pi[B_K(f_0,\varepsilon_n)] \asymp \varepsilon_n^k,\ k>0.$$

So the ratio is at most $C\varepsilon_n^m\, e^{2n\varepsilon_n^2} \to \infty$.

Nonparametric setting ($f \in L^2[0,1]$, infinite-dimensional): prior masses can be of exponential order in $n\varepsilon_n^2$. Typically:

$$\Pi[B_K(f_0,\varepsilon_n)] \sim e^{-C\, n\varepsilon_n^2}, \qquad \Pi[A_n] \sim e^{-C'\, n\varepsilon_n^2}.$$

Now both sides are exponential, and the condition becomes a delicate balance between the exponents C and C'. This is precisely what makes Bayesian nonparametrics nontrivial.

For a set Θ_n and $\varepsilon > 0$, the covering number $N(\varepsilon, \Theta_n, \|\cdot\|_n)$ is the minimal number of $\|\cdot\|_n$-balls of radius ε needed to cover Θ_n.
Its logarithm $\log N(\varepsilon, \Theta_n, \|\cdot\|_n)$ is the metric entropy: it measures the "effective dimension" of Θ_n at resolution ε.

Theorem [rate control with entropy]

Let ε_n be such that $n\varepsilon_n^2 \to \infty$ as $n \to \infty$. Assume there exist $C, D > 0$ and sets Θ_n such that:

i) $\log N(\varepsilon_n, \Theta_n, \|\cdot\|_n) \leqslant Dn\varepsilon_n^2$ (entropy)

ii) $\Pi[\Theta_n^c] \leqslant e^{-n\varepsilon_n^2(C+4)}$ (prior complement)

iii) $\Pi[B_K(f_0, \varepsilon_n)] \geqslant e^{-Cn\varepsilon_n^2}$ (prior mass near f_0)

Then for M large enough, the posterior contracts at rate $M\varepsilon_n$: as $n \to \infty$,

$$E_{f_0}\, \Pi\big[\|f - f_0\|_n \geqslant M\varepsilon_n \,|\, Y\big] = o(1).$$

Interpreting the conditions

The three conditions each have a clear role:

i) Entropy condition: $\log N(\bar{\varepsilon}_n, \Theta_n, \|\cdot\|_n) \leqslant Dn\varepsilon_n^2$
Controls the complexity of the sieve Θ_n: limits the number of "essentially different" regression functions at resolution ε_n.

ii) Prior mass on the complement: $\Pi[\Theta_n^c] \leqslant e^{-n\varepsilon_n^2(C+4)}$
The prior puts exponentially small mass outside the sieve, ensuring the prior doesn't place too much mass on overly complex functions.

iii) Prior mass near f_0: $\Pi[B_K(f_0, \varepsilon_n)] \geqslant e^{-Cn\varepsilon_n^2}$
The prior puts enough mass in a neighborhood of the true regression function, so we have a good prior.

Gaussian process priors for regression

GPs are natural candidates for prior distributions on regression functions on $[0,1]$.

Examples of GP priors:

Brownian motion (BM): (B_t) with covariance $K(s,t) = s \wedge t$. Sample paths are $1/2 - \varepsilon$ Hölder $\Rightarrow$ regularity $\approx 1/2$.

Riemann–Liouville process R^α: For $\alpha > 0$,

$$R_t^\alpha = \int_0^t (t-s)^{\alpha - 1/2}\, dB(s).$$

Has Hölder regularity close to α.

Gaussian series prior: For $\zeta_j \overset{iid}{\sim} \mathcal{N}(0,1)$ and (e_j) an ONB of $L^2[0,1]$,

$$W_t = \sum_{j=1}^{\infty} j^{-1/2-\alpha}\, \zeta_j\, e_j(t),$$

with regularity nearly α (in a Sobolev sense).

RKHS of a Gaussian process

Let $W = (W_t)_{t\in T}$ be a centered Gaussian process, with samples in $L^p[0,1]$, $p > 1$. Consider the first order chaos:

$$\mathcal{C}_W = \overline{\mathrm{Vect}\{W(t),\ t \in T\}}^{L^p} = \overline{\left\{\sum_{i=1}^{m} \alpha_i W(t_i),\quad t_i \in T,\ m \geqslant 1\right\}}^{L^p}.$$

RKHS

The Reproducing Kernel Hilbert Space of W is the set

$$\mathbb{H} = \left\{g_H : T \to \mathbb{R},\ \ g_H(t) = E[W_t H],\ \ H \in \mathcal{C}_W\right\},$$

equipped with the inner product: for $H_1, H_2 \in \mathbb{H}$,

$$\langle g_{H_1},\, g_{H_2}\rangle_{\mathbb{H}} = E[H_1 H_2].$$

The map $H \mapsto g_H$ from $\mathcal{C}_W$ into $\mathbb{H}$ is an isometry, so $\mathbb{H}$ is a Hilbert space. We can identify $\mathbb{H}$ with a subspace of $L^p[0,1]$.

RKHS: key properties

Reproducing property: For any $g_H \in \mathbb{H}$,

$$g_H(t) = E[W_t H] = \langle K(t,\cdot),\, g_H(\cdot)\rangle_{\mathbb{H}},$$

where $K(s,t) = E[W_s W_t]$ is the covariance function. This expresses the value of any RKHS element at a point via an inner product with K.

Example 1: Gaussian vector in $\mathbb{R}^k$. If $W \sim \mathcal{N}(0,\Sigma)$ with Σ invertible, then

$$(\mathbb{H}, \|\cdot\|_{\mathbb{H}}) = (\mathbb{R}^k, \|\cdot\|_{\mathbb{H}}), \qquad \langle u, v\rangle_{\mathbb{H}} = u^T \Sigma^{-1} v.$$

The RKHS is $\mathbb{R}^k$ itself, but with a twisted geometry given by Σ^{-1}.

Example 2: Random series. If $W_t = \sum_{j=1}^{\infty} \sigma_j \zeta_j e_j(t)$, $\zeta_j \overset{iid}{\sim} \mathcal{N}(0,1)$, (e_j) an ONB of $L^2[0,1]$, then

$$\mathbb{H} = \left\{ \sum_{j=1}^{\infty} \lambda_j e_j(t) \;:\; \sum_{j=1}^{\infty} \sigma_j^{-2} \lambda_j^2 < \infty \right\}, \qquad \left\langle \sum_j \lambda_j e_j, \sum_j \mu_j e_j \right\rangle_{\mathbb{H}} = \sum_{j=1}^{\infty} \sigma_j^{-2} \lambda_j \mu_j.$$

Smaller $\sigma_j \Rightarrow$ harder to include high-frequency components in $\mathbb{H}$.

Concentration function

Let W be a Gaussian process with sample paths in $L^p[0,1]$, with RKHS $\mathbb{H}$. For any $\varepsilon > 0$ and $w \in \overline{\mathbb{H}}^{L^p[0,1]}$, define

$$\varphi_w(\varepsilon) = \inf_{h \in \mathbb{H},\ \|h-w\|_{L^p} < \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log P[\|W\|_{\mathbb{B}} < \varepsilon]$$

$$=: \varphi_w^A(\varepsilon) + \varphi_0(\varepsilon).$$

The function $\varphi_w(\cdot)$ is called the concentration function of the process W.

$\varphi_w^A(\varepsilon)$: approximation term — how well the true regression function w can be approximated by elements of $\mathbb{H}$

$\varphi_0(\varepsilon) = -\log P[\|W\|_{L^p} < \varepsilon]$: small ball probability term — intrinsic complexity of the GP

Pre-concentration theorem

The concentration function allows us to verify the concentration conditions directly.

Theorem

Let W be a Gaussian process as above, with RKHS $\mathbb{H}$. Let $w_0 \in \overline{\mathbb{H}}^{L^p[0,1]}$ and let $\varepsilon_n > 0$ be such that

$$\varphi_{w_0}(\varepsilon_n) \leqslant n\varepsilon_n^2.$$

Then for any $C > 1$ with $Cn\varepsilon_n^2 > \log 2$, there exists $B_n \subset L^p[0,1]$ measurable such that

(i) $\log N(3\varepsilon_n, B_n, \|\cdot\|_{L^p}) \leqslant 6Cn\varepsilon_n^2$ ← entropy

(ii) $P[W \notin B_n] \leqslant e^{-Cn\varepsilon_n^2}$ ← prior complement

(iii) $P[\|W - w_0\|_{L^p} < 2\varepsilon_n] \geqslant e^{-n\varepsilon_n^2}$ ← prior mass

These are exactly the three conditions of the previous Theorem, with $\|\cdot\|_{L^p}$ instead of $\|\cdot\|_n$! But $\|\cdot\|_n \leqslant \|\cdot\|_{L^\infty}$, so the posterior contraction rate in regression is determined by solving $\varphi_{w_0}(\varepsilon_n) \leqslant n\varepsilon_n^2$.

Posterior convergence for GP priors in regression

Theorem

Let $Y_i = f(t_i) + \varepsilon_i$, $i = 1, \ldots, n$, be observations from the nonparametric regression model. Let Π be a prior on f, defined as the distribution of a centered Gaussian process in $L^\infty[0,1]$, with RKHS $\mathbb{H}$. Suppose the true regression function $f_0 \in \overline{\mathbb{H}}^{L^\infty[0,1]}$ and let ε_n satisfy

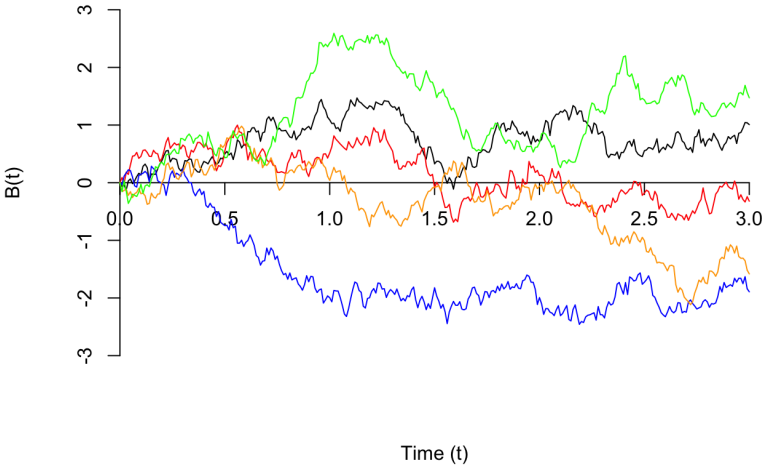
$$\varphi_{f_0}(\varepsilon_n) \leqslant n\varepsilon_n^2,$$

where φ_{f_0} is the concentration function of W in $L^\infty[0,1]$. Then for M large enough, as $n \to \infty$,

$$E_{f_0}\, \Pi\big[\|f - f_0\|_n > M\varepsilon_n \,|\, Y\big] \to 0.$$

Recipe: to find the posterior contraction rate for a GP prior in regression, one only needs to solve the concentration function inequality $\varphi_{f_0}(\varepsilon_n) \leqslant n\varepsilon_n^2$.

Brownian motion



Example: Brownian motion prior in regression

Consider the GP prior B_t (Brownian motion) in $L^\infty[0,1]$.

Covariance kernel: $K(s,t) = E[B_s B_t] = s \wedge t$. Since $K(s,\cdot)$ is piecewise linear, the RKHS elements are obtained by "smoothly combining" such functions.

RKHS: $\mathbb{H} = \left\{ \int_0^\cdot g(u)\,du,\ g \in L^2[0,1] \right\}$, with $\|h\|_{\mathbb{H}}^2 = \int_0^1 g(u)^2\,du = \|h'\|_2^2$.
Elements of $\mathbb{H}$ are absolutely continuous with square-integrable derivative and $h(0) = 0$.

Small ball probability: As $\varepsilon \to 0$,

$$\varphi_0(\varepsilon) = -\log P[\|B\|_\infty < \varepsilon] \asymp \varepsilon^{-2}.$$

Lemma
Suppose $w_0 \in C^\beta[0,1]$, for some $\beta \geqslant 0$ and $w_0(0) = 0$. Then

$$\inf_{h \in \mathbb{H}:\ \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{\frac{2\beta - 2}{\beta}} \vee 1.$$

Regression rates for the Brownian motion prior

Combining the previous results, with $a \vee b = \max(a, b)$:

$$\varphi_{w_0}(\varepsilon_n) \lesssim \underbrace{\varepsilon_n^{(2\beta-2)/\beta} \vee 1}_{\text{approximation}} + \underbrace{\varepsilon_n^{-2}}_{\text{small ball}}.$$

Equating to $n\varepsilon_n^2$ gives:

$$\varepsilon_n \asymp n^{-1/4} \vee n^{-\beta/2} = n^{-\{\frac{1}{4} \wedge \frac{\beta}{2}\}}.$$

Fastest rate: at $\beta = 1/2$, giving $\varepsilon_n \asymp n^{-1/4}$.

If $\beta < 1/2$: rate is $n^{-\beta/2}$ — the approximation term ("bias") dominates.

If $\beta \geqslant 1/2$: rate is $n^{-1/4}$ — the small ball probability ("variance") dominates.

Minimax comparison: The minimax rate for estimating C^β regression functions is $n^{-\beta/(2\beta+1)}$. The BM prior achieves this if and only if $\beta = 1/2$.

Regression rates for other GP priors

Riemann–Liouville process $W_t = R_t^\alpha$: Same analysis as BM, with the prior regularity $1/2$ replaced by α. Up to logarithmic factors:

$$\varepsilon_n \asymp n^{-\frac{\alpha \wedge \beta}{2\alpha+1}}.$$

Matches the minimax regression rate $n^{-\beta/(2\beta+1)}$ if and only if $\alpha = \beta$.

GP series prior: $W_t = \sum_{j=1}^\infty j^{-1/2-\alpha} \zeta_j e_j(t)$, with (e_j) an ONB of $L^2[0,1]$. For $L^2[0,1]$, a rate solving $\varphi_{w_0}(\varepsilon_n) \lesssim n\varepsilon_n^2$ is again:

$$\varepsilon_n \asymp n^{-\frac{\alpha \wedge \beta}{2\alpha+1}}.$$

Common pattern: In all cases, the posterior contraction rate for estimating the regression function is optimal if and only if the prior regularity α matches the true regularity β of f_0.

Take-away message

When a Gaussian process is used as prior on the regression function, the posterior contraction rate is determined by solving

$$\varphi_{f_0}(\varepsilon_n) \lesssim n\varepsilon_n^2,$$

where φ_{f_0} is the concentration function of the GP prior.

The results apply to many GP priors used in regression: Brownian motion, Riemann–Liouville, GP series, squared-exponential, Matérn, ...

One always obtains posterior consistency (rate $\to 0$), but the rate is minimax-optimal if and only if the prior regularity α matches the true regularity β of f_0.

⚠Since β is rarely known in practice, one needs adaptation — a prior achieving the optimal regression rate regardless of β.

Final exam logistic

Final exam will be held in person on Monday, April 13, at 2-5PM Toronto local time in room IA2010.

Exam will be 100 points in total and 180 mins long. Students are required to be at the exam location at least 10 mins early, with valid identification.

⚠No aid sheet allowed this time! A recap table on probability distributions and relevant functions will be distributed during the exam.

Except for Lecture 8, you should know the content of the theorems and be able to reproduce the computations we have done together in class.

Exam covers all lectures **except** the last one from today, it is closed book/internet.

OH will be held next week: Thursday, 3-6PM / Practice exam will soon be shared!

Closing remarks

We focused in the role of Bayesian methods in modern statistical inference:

- The posterior is a rich object: position (mean, median) + uncertainty (variance, credible regions)
- Decision theory provides principled estimators adapted to your loss function
- Bayesian and frequentist approaches agree asymptotically (Bernstein–von Mises)
- The choice of prior matters less as n grows, but remains crucial in high/infinite dimension
- MCMC/VI made Bayesian statistics practical: only need the posterior up to a constant
- Next steps: more general nonparametric Bayes, prior elicitation,...

Study for the final!

Review lectures.

Understand derivations (try to reproduce without help).

Solve the exercises and midterm.

Fill out course evaluations!

Most importantly: Good luck!