Гауссовские процессы

Created by	 Yana Savchenko
Telegram	https://t.me/before_relu_or_after

В данной заметке расскажем вам про то, что такое гауссовские процессы и про их применение для регрессии и для байесовской оптимизации.

1. Случайные процессы и гауссовские процессы

В мире много динамических процессов, которые содержат в себе некоторую неопределенность:

- Курс акций;
- Функция потерь в машинном обучении;
- Положение какой-то частицы нанометрового размера во взвеси (коллоидные системы);
- Положение клиента, который перемещается внутри магазина от прилавка к прилавку.

Проблемы со всеми этими вещами в том, что мы, даже зная всю историю, чаще всего не можем спрогнозировать поведение в таких системах. Но как-то работать с этим всем добром хочется, хочется уметь оценивать вероятности и что-то прогнозировать. Для этого нам нужна модель.

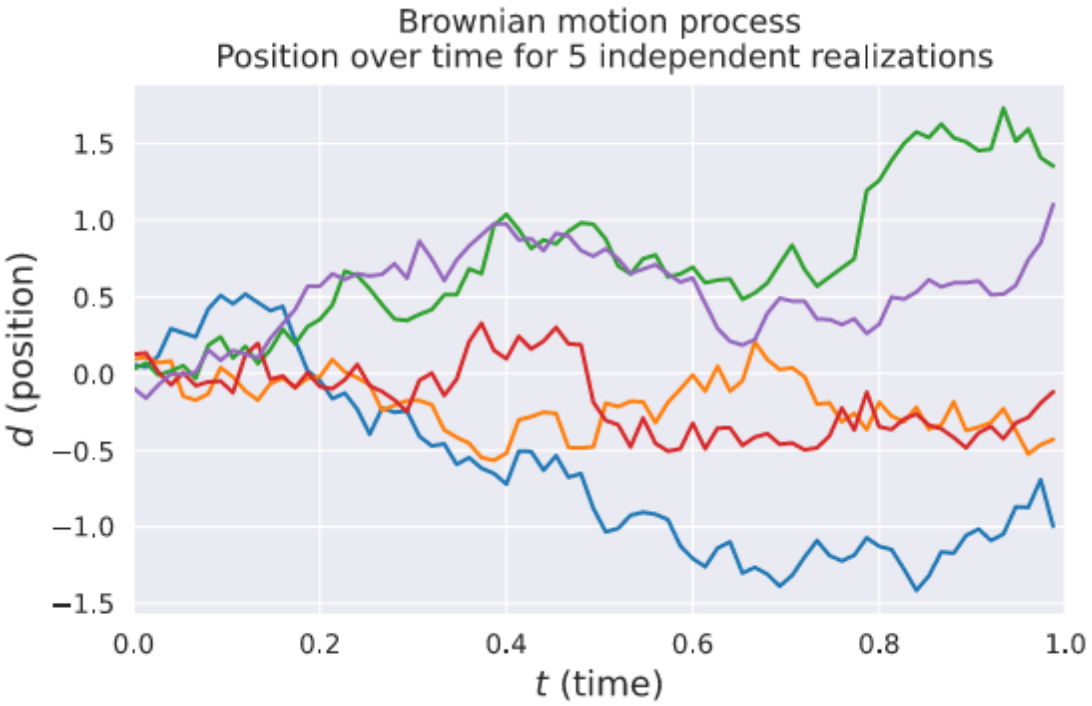
Описанные выше динамические процессы мы будем называть случайными процессами $\xi(\omega, t)$. Обычную случайную величину мы могли бы обозначить $\xi(\omega)$ — это функция исхода (например, исходами являются выпадение орла или решки при подбрасывании монеты). Чтобы перейти от стационарных процессов к динамическим, мы добавили в аргумент время t и получили $\xi(\omega, t)$. Роль динамической переменной t может выполнять не только время,

но и, например, координата в пространстве. Переменная t может быть дискретной или непрерывной, можно воспринимать её просто как индекс:

- Например, если мы будем бросать монетку каждый день, то зададим случайный процесс с дискретным временем.
- Пример случайного процесса с непрерывным временем — это цены на акции. Для разных моментов времени t в случайном процессе случайные величины могут быть как независимыми (пример с монетками), так зависимыми (цены на акции).

А теперь посмотрим, что будет, если зафиксировать аргументы $\xi(\omega, t)$:

- Если мы зафиксируем время $t = t_0$ в случайном процессе, то получим обычную случайную величину $\xi(\omega, t_0)$ (сечение процесса). Например, если наша монетка ржавеет, то для каждого t у нас будет немного разная вероятность получить орла (ржавчина налипает, и центр тяжести монетки смещается), то есть у каждого сечения будет разная функция распределения.
- Если мы зафиксируем исход $\omega = \omega_0$, то получим обычную функцию $\xi(\omega_0, t)$ (реализация, траектория процесса) — например, какой-нибудь синус или косинус. Сложность в том, что предыстория процесса может быть одинаковой для разных исходов (представьте себе два графика с курсом акций, которые до какого-то дня идеально совпадали, а потом разошлись). И, глядя на одинаковый участок, мы не можем знать, какую реализацию из двух мы наблюдаем.
  Вот так, например, выглядят разные реализации $\xi(\omega_0, t)$ для одномерного броуновского движения:



Броуновское движение, известное со школы — тоже случайный процесс. t — это динамическая переменная случайного процесса, а полученные кривые — это разные реализации процесса $\xi(\omega_0, t)$

Ну и введём ещё пару понятий, которые связаны со случайными процессами:

1. Мат. ожидание случайного процесса $m_\xi(t)$ — это функция времени, которая в каждый момент времени t равна мат. ожиданию сечения случайного процесса;
2. Дисперсия случайного процесса $D_\xi(t)$ — это функция времени, которая в каждый момент времени t равна дисперсии сечения случайного процесса;
3. Ковариационная функция случайного процесса $k_\xi(t, s)$ — это функция времени t, s, которая равна мат. ожиданию произведения двух сечений $\mathbb{E}(\xi(t), \xi(s))$

Гауссовский процесс — это такой случайный процесс f, у которого сечения образуют вектор с нормальным распределением

$(\xi(\omega, t_1), \xi(\omega, t_2), \ldots, \xi(\omega, t_n))$, или, другими словами, $f(t) \sim N(m_\xi(t), k_\xi(t, t))$.

В дальнейшем индексы ξ будем для краткости опускать.

2. Немного математики: байесовский подход к вероятности

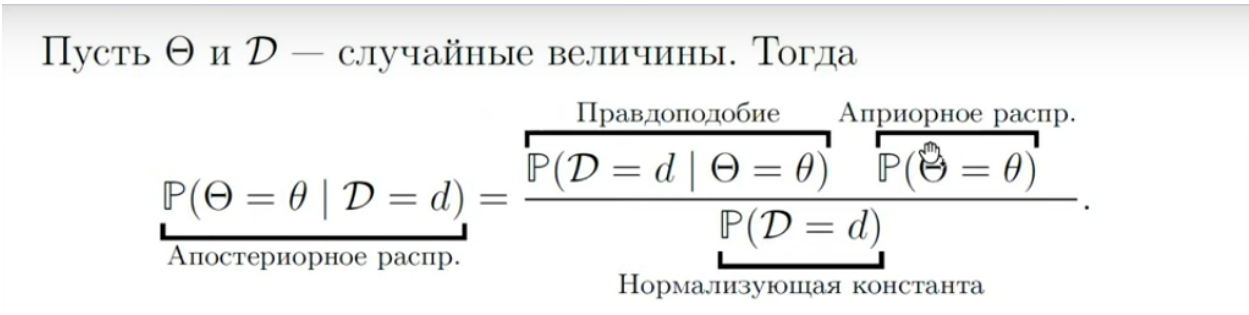
Так как дальше мы будем разбираться с байесовской оптимизацией, а также часто употреблять слова “априорное распределение” и “апостериорное распределение”, то давайте поговорим про байесовский подход к вероятности.

Давайте на примере задачи с подбрасыванием монетки разберёмся с тем, что такое байесовский подход к вероятности. Давайте обозначим “1” событие “выпал орёл” и “0” событие “выпала решка”. Обозначим за θ вероятность выпадения орла, тогда вероятность выпадения решки составит $1 - \theta$. А теперь мы подбросим монетку несколько раз и получим в результате последовательность d, например, вот такую: $d = 111011000111$. У нашей монетки в такой постановке задачи есть один параметр θ (вероятность получить орла), и мы хотим получить оценку этого параметра θ при помощи данных, которые есть у нас на руках.

Пользуясь частотным подходом (frequentist approach), которому нас всех учат в университете (🙂), мы можем получить, что вероятность выпадения орла — это *число*, при котором достигает максимума вероятность получить всю последовательность 111011000111:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(\theta^{\#1}(1-\theta)^{\#0}) = \underset{\theta}{\operatorname{argmax}}(\theta^8(1-\theta)^4)$$

Пользуясь байесовским подходом к вероятности (bayesian approach), мы получаем *не число, а распределение.* Мы трактуем вероятность выпадения орла — вполне себе детерминистическую величину — как некую случайную величину с каким-то распределением. Вспомним формулу Байеса:

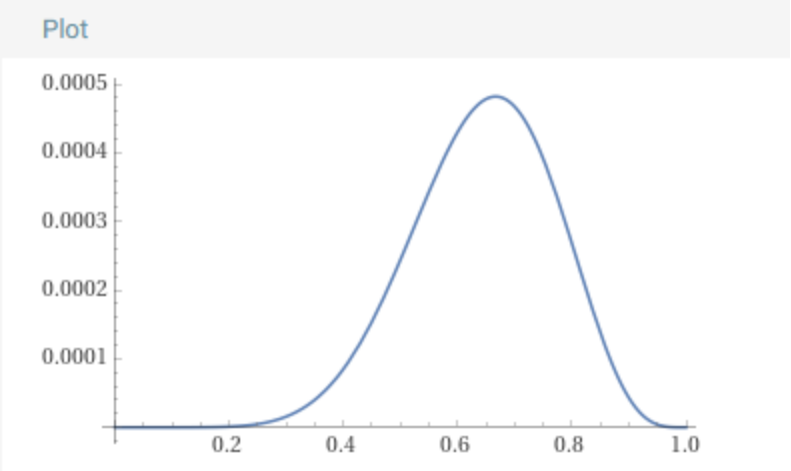


Мы хотим получить *апостериорное распределение вероятности* (posterior) для параметра монетки Θ (в нашем случае у монетки есть только один параметр — это вероятность выпадения орла) при условии, что мы наблюдали последовательность $d = 111011000111$. Эта величина обозначена как $P(\Theta = \theta | D = d)$, и она пропорциональна произведению двух величин: вероятности $P(D = d | \Theta = \theta)$ получить последовательность d в зависимости от θ (т.е. правдоподобия, или likelihood-а) и априорного распределения (prior-а) $P(\Theta = \theta)$ для вероятности выпадения орла θ. Априорное распределение мы должны задать сами.

Пример: предположим, что мы ничего не знаем о монетке, поэтому выберем равномерное априорное распределение $P(\Theta = \theta)$ для нашего параметра θ — вероятности получить орла. То есть мы считаем, что вероятность выпадения орла может принимать любое значение в диапазоне от 0 до 1 с одинаковой вероятностью. Тогда байесовский подход говорит нам о том, что апостериорное распределение пропорционально следующей величине:

$\hat{\theta} = P(\Theta = \theta | D = 111011000111) \propto \theta^{\#1}(1-\theta)^{\#0} \times 1 = \theta^8(1-\theta)^4 \times 1$
(единичка обозначает плотность равномерного распределения).

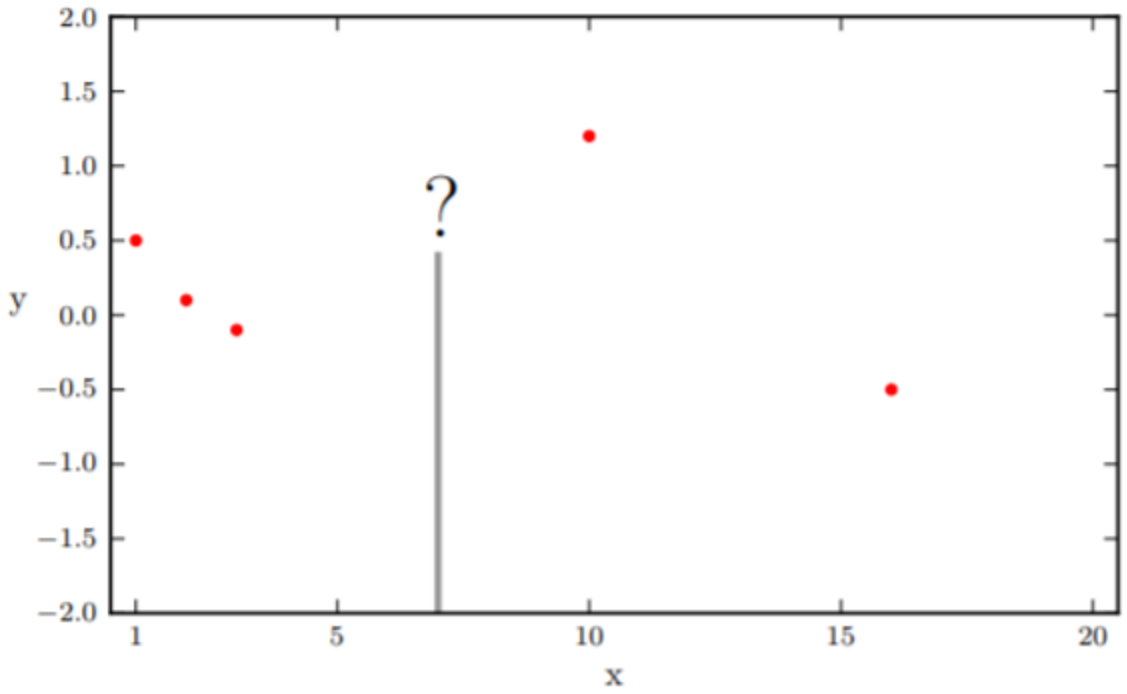
А вот так выглядит график $\hat{\theta} = \hat{\theta}(\theta)$:



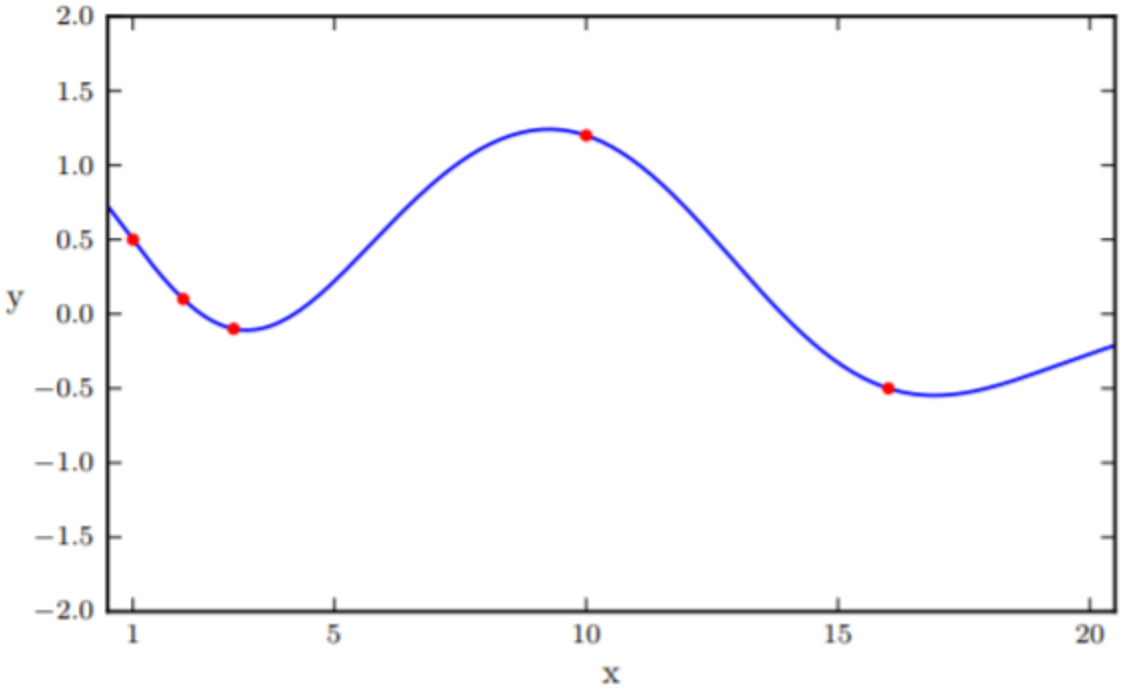
Ну и оценку для вероятности орла можно получить, как $argmax$ этого распределения $\hat{\theta}(\theta)$.

3. Гауссовские процессы и регрессия

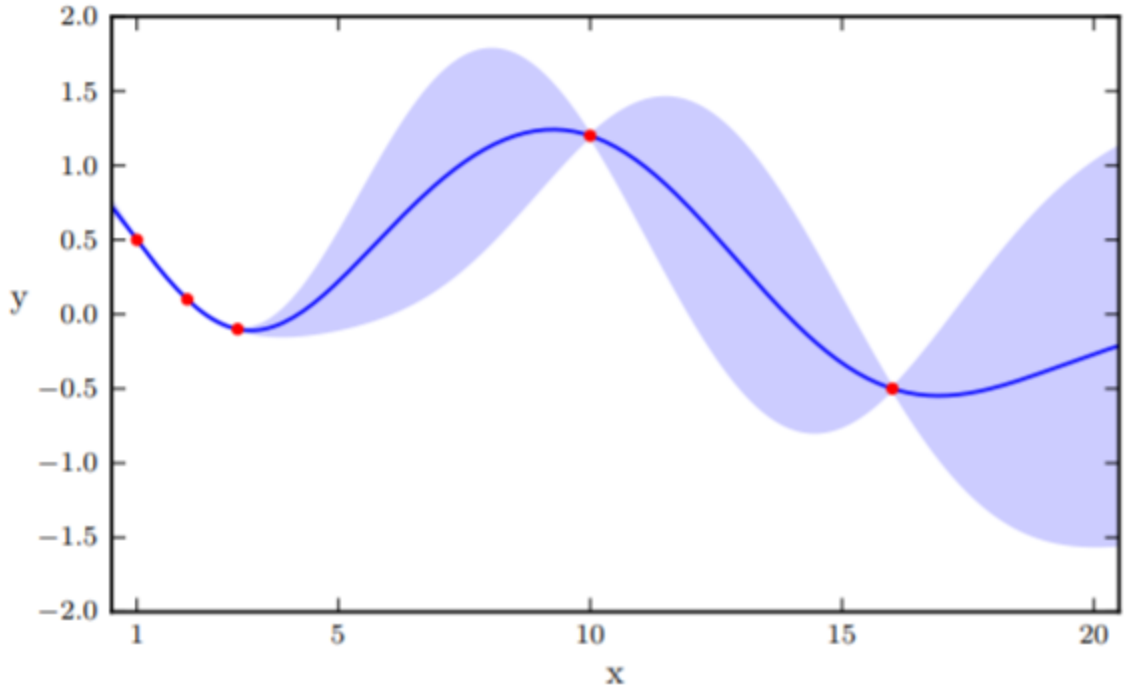
Представим, что у вас есть набор данных, который представляет собой множество пар точек (x, y) (обозначены красными точками на графике ниже). Мы хотим по этой ограниченной выборке попробовать предсказать, чему будет равно значение y для какой-то новой точки x, то есть решить задачу нелинейной регрессии.



У этой задачи в общем случае нет единственного решения. На практике, чтобы получить ответ, вам надо будет выбрать аппроксимирующую функцию (например, какой-нибудь полином достаточно высокой степени) и воспользоваться стандартным алгоритмом для нелинейной регрессии, чтобы определить параметры функции (полинома). В результате вы получите *одну функцию*, которая проходит через красные точки:



А на самом деле решение такой задачи — это не одна функция, а *семейство функций.* Например, таких, которые лежат внутри закрашенной голубым области:

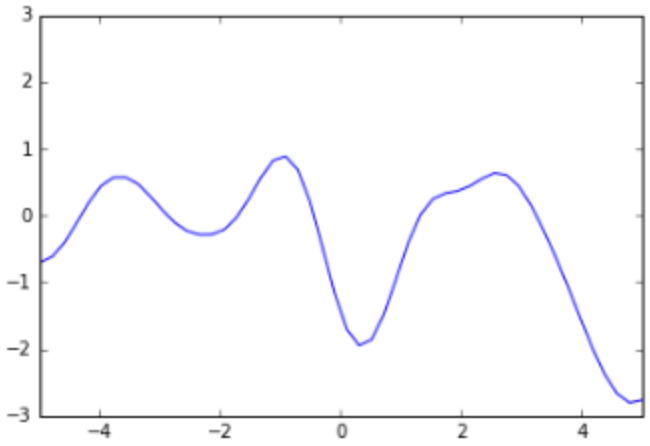


Возникает вопрос: а как поменялось бы наше решение, если бы в исходной выборке была ещё одна точка? Очень сильно или нет? В некоторых случаях нам бы очень хотелось получать не только зависимость, но ещё и оценку доверительного интервала. 🙂

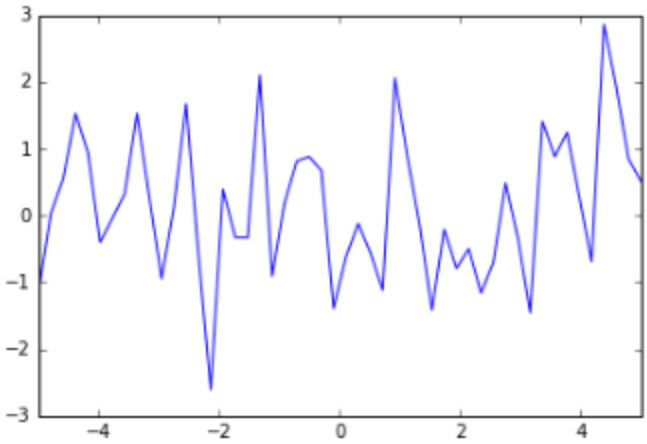
Давайте попробуем решить задачу регрессии при помощи гауссовских процессов. Будем считать, что точки, которые мы получили, были насемплированы из гауссовского процесса $f(x)$, где x выполняет роль динамической переменной. У гауссовского процесса $f(x)$ есть сечения $f(x_i)$ в каждой точке. Вот прям из этого сечения семплируем случайные величины и получаем наши красные точки $y(x_i)$.

Почему вообще мы считаем, что по этим точкам мы можем задать случайный процесс? Как следует из нашего определения, если мы знаем $\xi(\omega, t)$, то мы о нашем случайном процессе знаем всё. Оказывается, есть ещё один способ задать случайный процесс, который более удобен для практических задач. Как у случайной величины мы можем задать функцию распределения, так и для случайного процесса мы можем сделать нечто подобное, а именно: взять и задать функции распределения для всех точек в выборке. Это позволяет нам получить случайный процесс для наших точек, а может даже и не один (см. Теорема Колмогорова).

Обычно, чтобы решить задачу нелинейной регрессии, мы выбираем аппроксимирующую функцию (например, полином $n-$й степени) и ищем параметры функции с помощью какого-нибудь алгоритма. Сейчас у нас будет другой подход, непараметрический: мы будем считать, что значение нашей функции $f(x)$ в каждой точке x — это случайная величина с нормальным распределением $f(x) \sim N(m(x), k(x, x))$, где $m(x)$ и $k(x, x)$ — какие-то функции. Для тех, кто уже забыл: это прям определение гауссовского случайного процесса с функцией среднего $m(x)$ и ковариационной функцией (ядром) $k(x, x')$. На практике для гауссовского процесса обычно выбирают $m(x) = 0$. Ещё сразу определяются, какой вид должна иметь ковариационная функция. Ковариационные функции бывают разные, и от их вида очень сильно зависит решение:



Вот так может выглядеть результат семплирования из гауссовского процесса с $m(x) = 0$ и гладкой ковариационной функцией (или ядром) $k(x, x')$

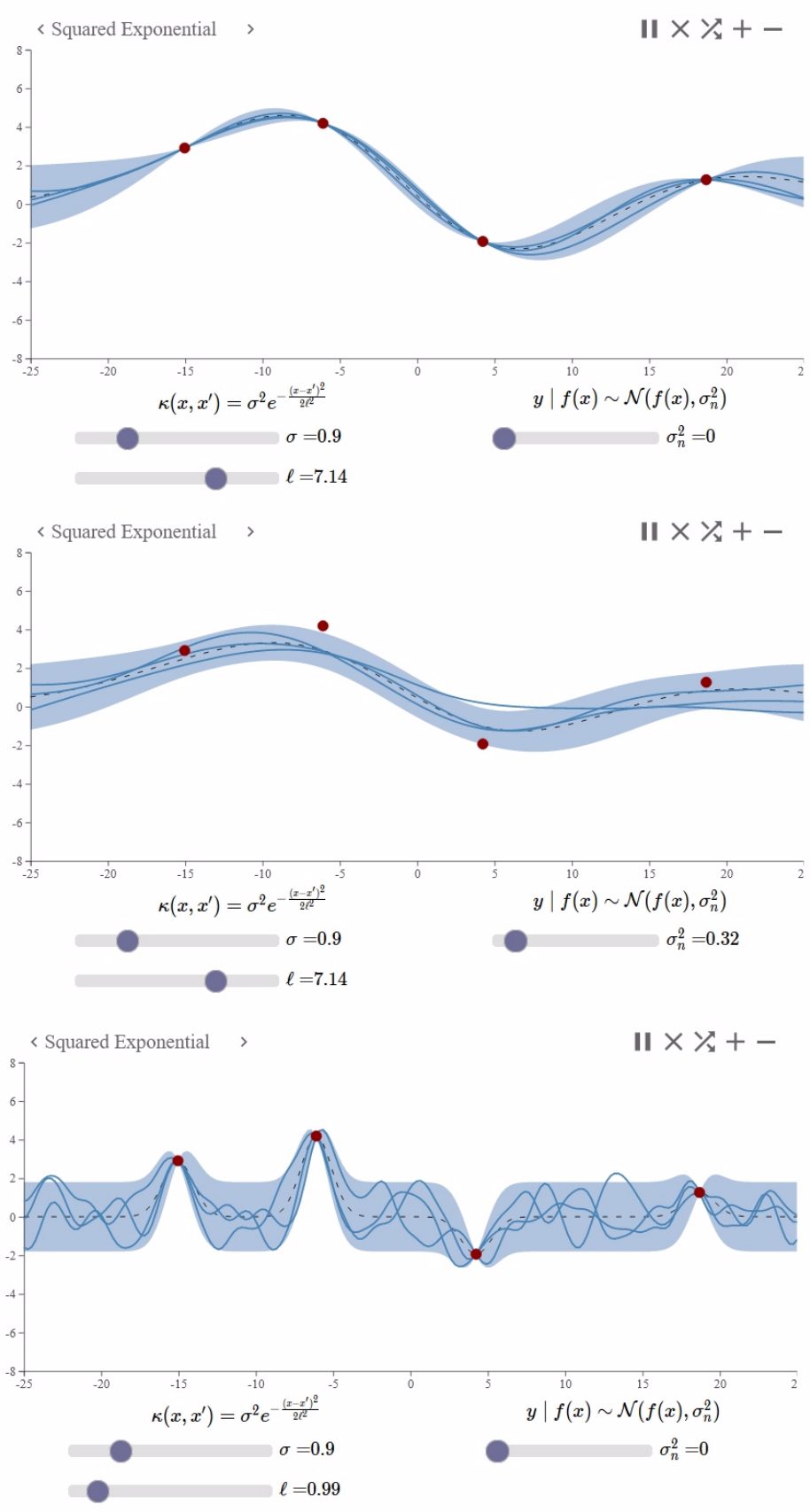


А если $k(x, x')$ — не гладкая, то результат семплирования может быть таким

▼ ***Лирическое отступление про разные ковариационные функции***

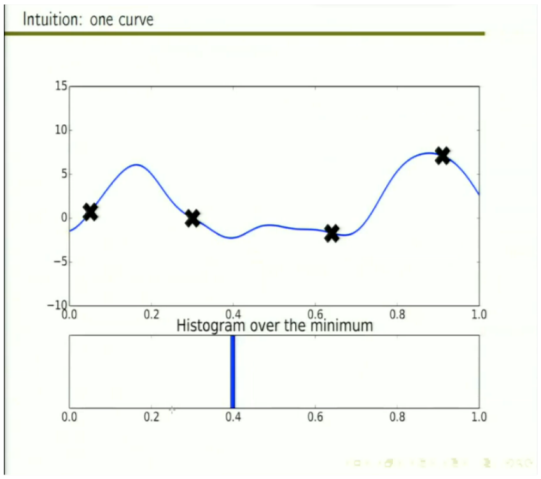
Можете попробовать вот тут поиграться с гауссовскими процессами: накиньте несколько точек, выберите ковариационную функцию. Вы получите множество кривых, каждая из которых проходит через ваши точки. А затем поменяйте параметры у ковариационной функции, чтобы $k(x, x')$ становилась больше (меньше) для соседних точек. Вы увидите, как увеличивается (уменьшается) доверительный интервал. Каждая из кривых строится не аналитически — чтобы построить её, используется генератор случайных чисел, который генерирует точки по заданному вами правилу. Ещё там же можно отрегулировать уровень шума (обозначен как $y|f(x) \sim N(f(x), \sigma_n^2)$) — то есть позволить подбирать

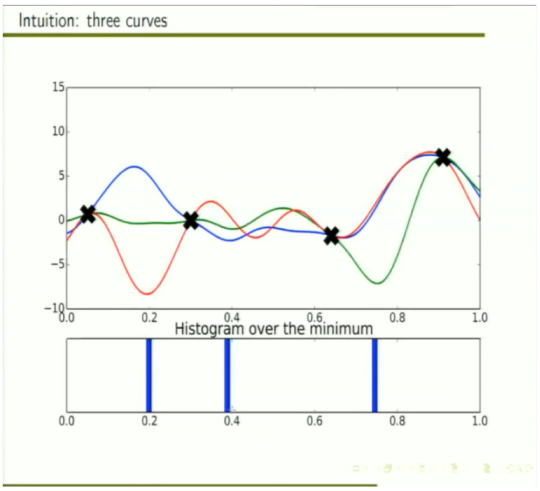
функции, не точь-в-точь проходящие по точкам выборки, а с небольшим отклонением.

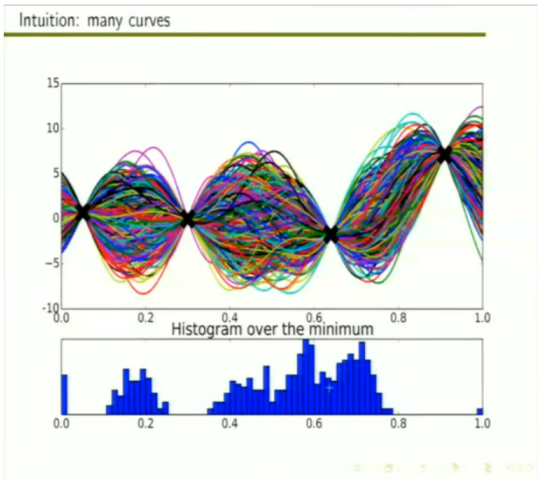

Squared Exponential
$\kappa(x, x') = \sigma^2 e^{-\frac{(x-x')^2}{2\ell^2}}$
$y \mid f(x) \sim \mathcal{N}(f(x), \sigma_n^2)$
σ =0.9
ℓ =7.14
σ_n^2 =0
Squared Exponential
$\kappa(x, x') = \sigma^2 e^{-\frac{(x-x')^2}{2\ell^2}}$
$y \mid f(x) \sim \mathcal{N}(f(x), \sigma_n^2)$
σ =0.9
ℓ =7.14
σ_n^2 =0.32
Squared Exponential
$\kappa(x, x') = \sigma^2 e^{-\frac{(x-x')^2}{2\ell^2}}$
$y \mid f(x) \sim \mathcal{N}(f(x), \sigma_n^2)$
σ =0.9
ℓ =0.99
σ_n^2 =0

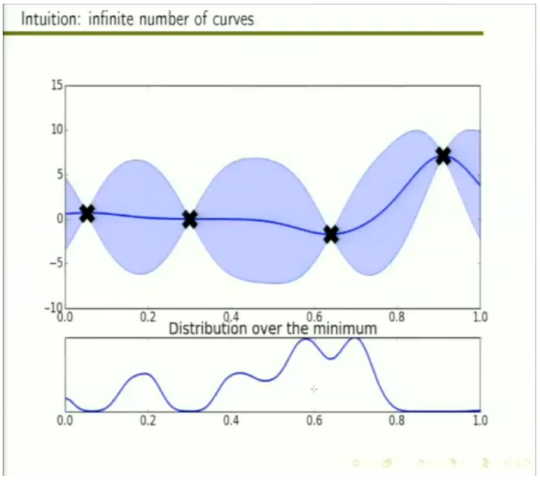
Гауссовский процесс по 4 точкам. Как различается результат, если немного поменять уровень шума или параметры ковариационной функции (ядра)

А теперь посмотрим на иллюстрации, как получается доверительный промежуток: сначала из гауссовского процесса по нашим точкам мы можем получить функцию. Затем вторую, третью. Семплируем ещё, и в пределе у нас как раз и получается что-то типа доверительного промежутка:









Про байесовский вывод гауссовских процессов

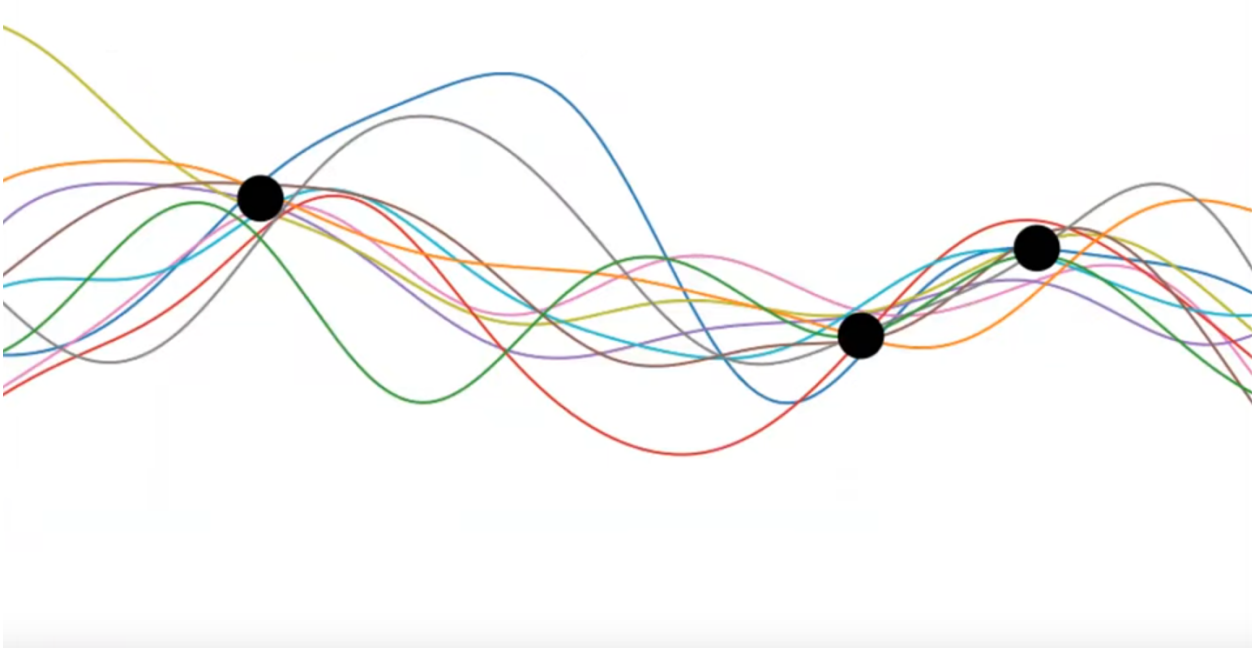
Вспомним ещё раз нашу любимую формулу Байеса:

$$P(\Theta = \theta | D = d) = \frac{P(D=d|\Theta=\theta) \times P(\Theta=\theta)}{P(D=d)},$$

где
$P(\Theta = \theta)$ — это априорное распределение, а $P(D = d|\Theta = \theta)$ — правдоподобие. Давайте будем считать, что априорное распределение $P(\Theta = \theta)$ — это траектории из гауссовского процесса, например, такие:

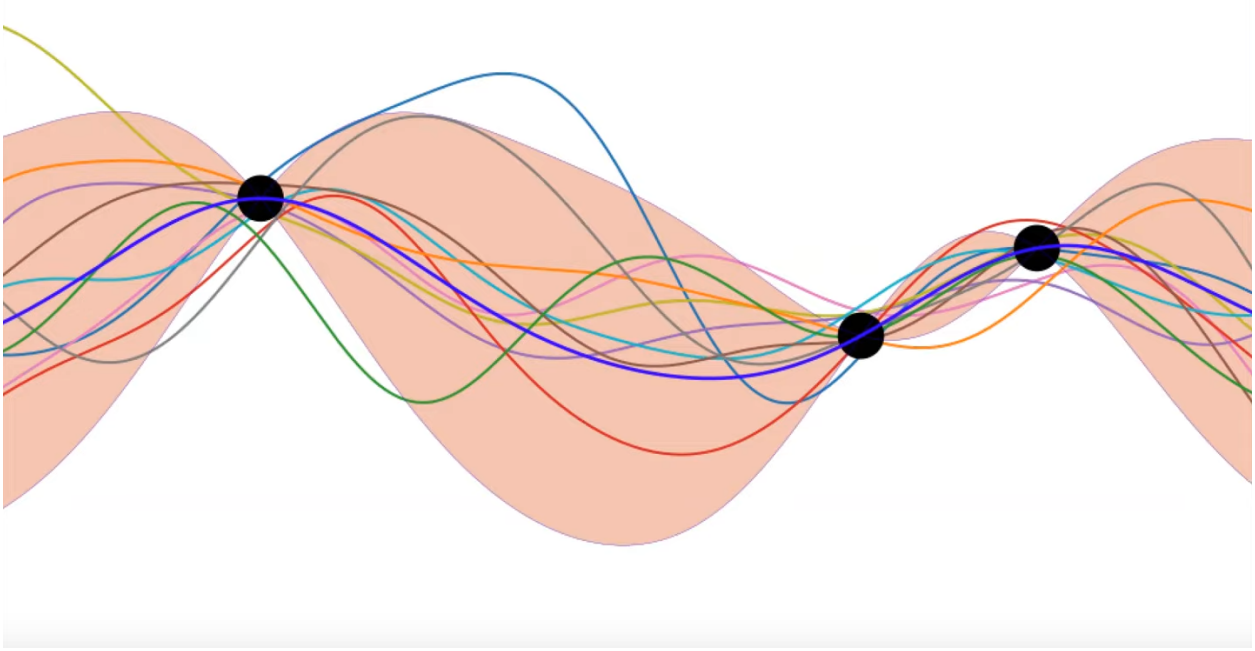
Априорный гауссовский процесс

А ещё у нас есть какие-то данные (точки). Если мы возьмем в качестве правдоподобия $P(D = d|\Theta = \theta)$ нормальное распределение, то у нас снова получится гауссовский процесс, только траектории уже будут проходить не как угодно, а по точкам:



Апостериорный гауссовский процесс

По параметрам получившегося апостериорного гауссовского процесса (который проходит по точкам) мы можем определить доверительный интервал. Можно для простоты считать, что мы насемплируем много-много траекторий из этого гауссовского процесса, а потом закрасим оранжевым такой промежуток, куда попадёт большинство траекторий:



Апостериорный гауссовский процесс + визуализация среднего и доверительного интервала

Алгоритм регрессии при помощи гауссовских процессов, или Как по данным $(x_1, y_1), \ldots, (x_n, y_n)$ получить разумную стохастическую модель, их интерполирующую:

1. Замерить значение y в нескольких точках x;
2. Взять с потолка априорное среднее $m(x)$ и параметрическое семейство априорных ковариационных функций $k_\theta(x, x')$. Обычно берут $m(x) = 0$.
3. По данным $(x_1, y_1), \ldots, (x_n, y_n)$ выбрать оптимальное значение параметров θ и шума σ^2 (метод максимального правдоподобия, какой-нибудь градиентный спуск)
4. По данным $(x_1, y_1), \ldots, (x_n, y_n)$ и априорным $m(x), k_\theta(x, x')$ получаем апостериорные $\tilde{m}, \tilde{k}$
5. Можем использовать $N(\tilde{m}(x), \tilde{k}(x, x))$ как стохастический прогноз в точке x.
6. Можем использовать функции из получившегося гауссовского процесса как возможные детерминистические модели (получать конкретные $y =$

$y(x)$, тип как с полиномом), просто семплируя значения в конкретных точках (см. пункт 5).

У регрессии при помощи гауссовских процессов есть следующие недостатки:

- Сложность построения моделей зависит от числа точек в выборке как $O(N^3)$. Это значит, что если у вас очень большая выборка точек, то вам, скорее всего, гауссовские процессы не подходят.
- Результат аппроксимации очень зависит от выбранной ковариационной функции. Самое распространенное ядро $k(x, x')$ — это обычная гауссова шапочка, и с ней получаются хорошие результаты, если надо аппроксимировать гладкую функцию. Чтобы подогнать негладкую функцию, например, функцию Хевисайда (ступеньку), лучше использовать какое-то другое ядро (например, экспоненциальное).
- Вырождение модели. Если точки выборки набросаны не равномерно по всему промежутку, а образовывают кластеры, тогда найденные аппроксимирующие функции будут содержать осцилляции и получаться менее гладкими. С этим можно бороться при помощи регуляризации.

Резюме: используем, если нужна оценка дисперсии и выборка не огромная

4. Байесовская оптимизация

Байесовская оптимизация — это итерационный метод, который позволяет определить минимум (или максимум) сложно вычислимой функции ϕ.
Пример функции ϕ:

- Лаборант Вася ставит эксперимент на физической установке. Каждый раз ему нужно выбирать параметры установки, а сам эксперимент занимает несколько дней (ϕ — это результат эксперимента в зависимости от параметров установки);
- Джун Коля хочет найти оптимальные гиперпараметры для обучения нейронной сети. Обучение сети от начала и до конца занимает день (ϕ — это метрика модели в зависимости от гиперпараметров обучения);

Обратим внимание, что нигде не требуется, чтобы ϕ была дифференцируемой.

Давайте для удобства будем считать, что мы ищем гиперпараметры, при которых наша нейронная сеть будет давать максимальную метрику. Пусть x — это гиперпараметры, а ω — это сид, фиксирующий случайность при обучении сети. Когда мы фиксируем гиперпараметры, то наша метрика является случайной величиной, которая зависит от сида. Когда мы фиксируем сид, то метрика является просто функцией гиперпараметров. Очень похоже на определение случайного процесса $\xi(\omega, x)$.

Обозначим $\phi(x) = \xi(\omega, x)$ метрику сети со знаком минус и будем итеративно искать *минимум* функции $\phi : \mathbb{R} \to \mathbb{R}^d$.

Алгоритм оптимизации следующий:

1. Вычисляем в значение функции n точках $x_1, .., x_n$:

   $\phi(x_1), \ldots, \phi(x_n)$

2. Делаем регрессию на этой выборке. Иными словами, получаем апостериорный гауссовский процесс f по нашим данным:

   $(x_1, \phi(x_1)), \ldots, (x_n, \phi(x_n))$

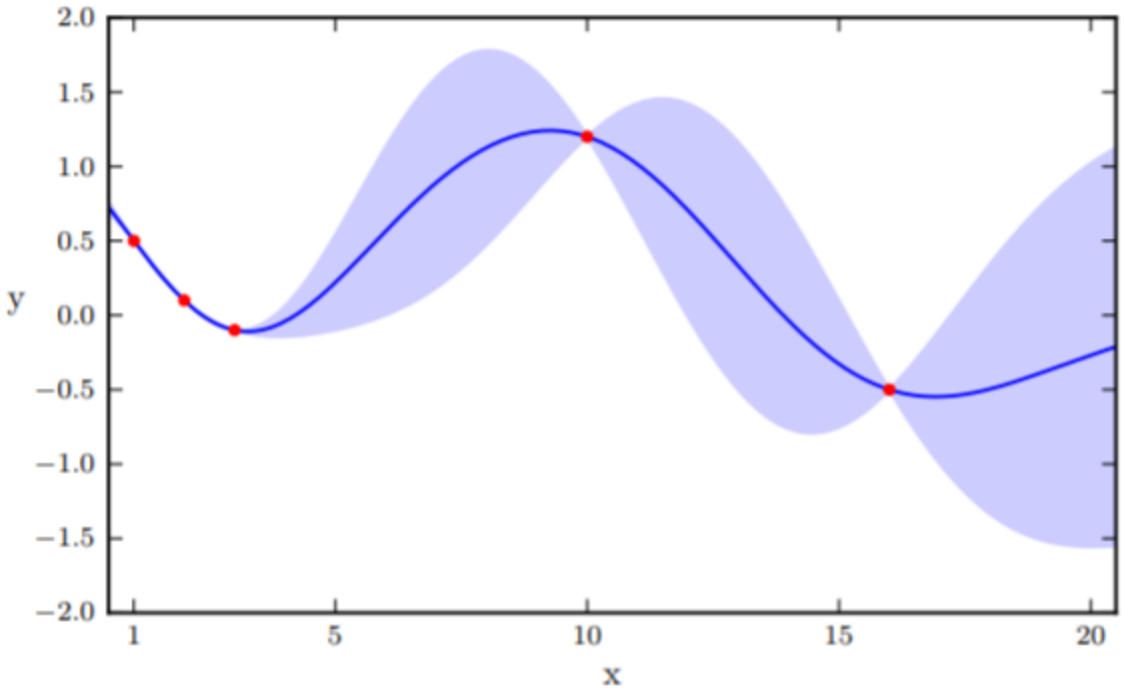
3. Выбираем следующую точку x_{n+1}, чтобы произвести в ней измерение. Для удобства переобозначим индексы у $x_1, .., x_n$ так, чтобы $\phi(x_1) \geq \phi(x_2) \geq ... \geq \phi(x_n)$. Правило, по которому мы выбираем x_{n+1}, будет следующим:

   $x_{n+1} = \underset{\mathbb{R}^d}{\operatorname{argmax}}(\alpha(x|S_n))$, где $\alpha(x|S_n)$ — это некая дифференцируемая функция, которая зависит от истории наблюдений $S_n = (x_1, \phi(x_1)), \ldots, (x_n, \phi(x_n))$. Эта функция говорит нам о том, в каких местах траектории лучше всего сделать следующий замер ϕ. Функция ϕ считается очень долго, поэтому нам не хотелось бы просто идти по координатной сетке и делать замеры: это очень дорого в плане вычислений, мы хотим выбирать точки для замеров по-умному.

Вычисление самой функции α и максимума от неё на фоне остальных дорогих вычислений ничего не стоит.

▼ Про $\alpha(x|S_n)$

$\alpha(x|S_n)$ называется *acquisition function.* Давайте подумаем, какой она может быть. Мы не можем просто взять наш гауссовский процесс и сказать, что x_{n+1} нужно искать в минимуме функции f: мы помним, что получить траекторию из случайного процесса по точкам можно многими способами. Так или иначе, выбор α должен выражать вероятностный характер траектории. Или, другими словами, мы должны помнить, что у нас есть не только траектория, но и доверительный интервал, который, вообще говоря, может быть большим:



Не забываем про доверительный интервал (заштрихованную голубым область)

Вот некоторые часто используемые $\alpha(x)$:

1. $P(f(x) < \phi(x_n))$ — Maximal Probability of Improvement.
   Выбираем
   x, максимизирующий вероятность того, что $f(x)$ будет меньше текущего минимума (напомню, мы хотим найти минимум функции).

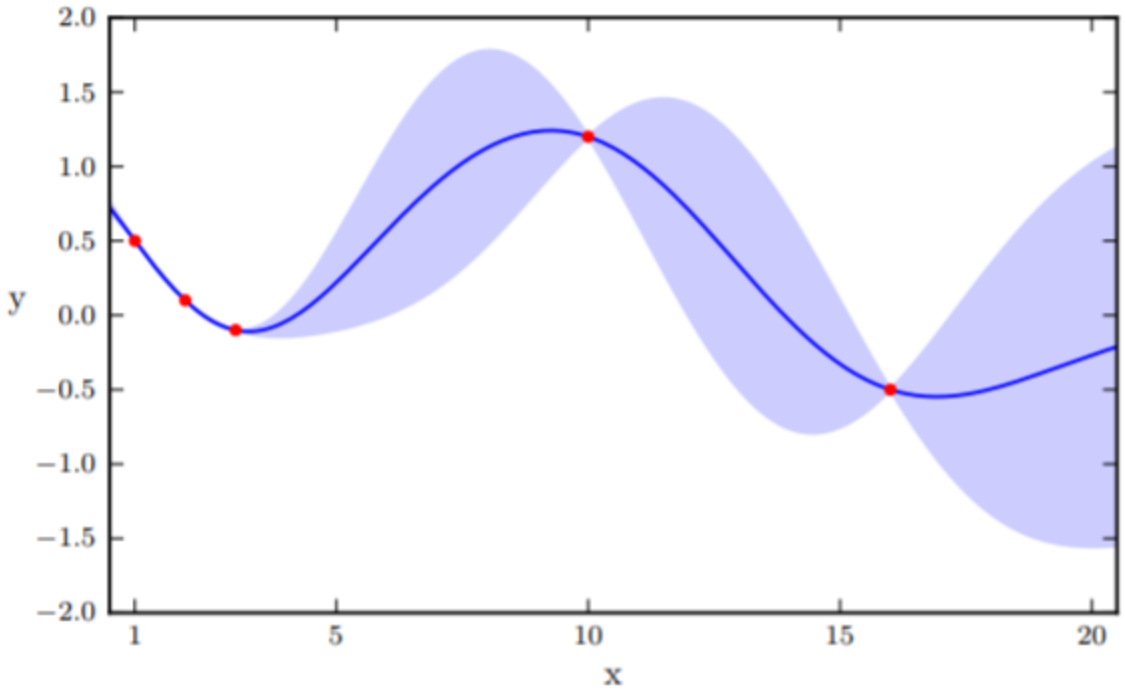
Так как у нас есть f, который мы построили по n измерениям, мы можем просто посчитать $argmax$;

2. $\mathbb{E}max(\phi(x_n) - f(x), 0)$ — Expected Improvement.
   Выбираем
   x, максимизирующий ожидаемое улучшение

4. Добавляем x_{n+1} в выборку, переходим к пункту 2, но уже с выборкой из $n + 1$ измерений.

5. Повторяем до тех пор, пока у нас больше нет времени или вычислительных ресурсов на оптимизацию. Из получившихся экспериментов выбираем лучший.

Решение таких задач — это своего рода поиск баланса между двумя тактиками: *exploration* и *exploitation*. Тактика *exploitation* приводит к тому, что мы ищем минимум там, где среднее у вероятностной модели f низкое (см. синюю линию). Тактика *exploration* говорит нам о том, что можно испытать удачу и поискать минимум в той области, где дисперсия большая (см. синий доверительный интервал):



Всё ещё не забываем про доверительный интервал (заштрихованную голубым область)

Резюме: используем для нахождения максимума/минимума трудно вычислимой функции (из-за времени или ресурсов), которая не обязана быть дифференцируемой. Работает хорошо при невысоких размерностях, так что лучше не оптимизировать много гиперпараметров.

Ссылки

- http://smlbook.org/GP/ — визуализация гауссовских процессов (интерактивная);
- https://www.youtube.com/watch?v=DU_Lhl_j2v0 — короткий доклад про гауссовские процессы;
- https://www.youtube.com/watch?v=0rxwLKv1ou0 — длинный, но классный доклад про гауссовские процессы в машинном обучении;
- https://academy.yandex.ru/handbook/ml/article/podbor-giperparametrov — про подбор гиперпараметров с помощью байесовской оптимизации;
- https://www.youtube.com/watch?v=9pOyslYQH50&list=PL4_hYwCyhAvYVM77cTV0WspsLbtASnAy4 — лекции МФТИ по случайным процессам (плейлист);
- https://www.youtube.com/watch?v=9BiUAVfOelA — доп. семинар МФТИ по гаусовским процессам и байесовской оптимизации в машинном обучении;