

# Материалы к лекции по вариационному подходу по спецкурсу «Байесовские методы машинного обучения» 2010

20 декабря 2010 г.

Ликбез

## 0.1. Дивергенция Кульбака-Лейблера.

Дивергенция Кульбака-Лейблера  $KL(q||p)$  является мерой расстояния между двумя вероятностными распределениями  $q(X)$  и  $p(X)$ :

$$KL(q||p) = - \int q(X) \log \frac{p(X)}{q(X)} dX.$$

КЛ-дивергенция обладает следующими свойствами:

- $KL(q||p) \geq 0$  и  $= 0 \Leftrightarrow q(X) \equiv p(X)$ ,
- $KL(q||p) \neq KL(p||q)$ .

Доказательство первого свойства.

Рассмотрим произвольную строго вогнутую функцию  $f(x)$ . Тогда по определению

$$f(\alpha x_1 + (1 - \alpha)x_2) \geq \alpha f(x_1) + (1 - \alpha)f(x_2), \quad \forall \alpha \in [0, 1].$$

В этом неравенстве равенство достигается, если  $x_1 = x_2$  или при  $\alpha = 0, 1$ . Данное неравенство может быть обобщено на случай большего числа точек  $x$ :

$$f(\alpha_1 x_1 + \dots + \alpha_N x_N) \geq \alpha_1 f(x_1) + \dots + \alpha_N f(x_N), \quad \sum_{n=1}^N \alpha_n = 1, \alpha_n \geq 0.$$

Это неравенство известно как неравенство Йенсена и может быть легко доказано по индукции. Переходя в этом неравенстве от сумм к интегралам, получаем следующее обобщение:

$$f\left(\int \alpha(T) X(T) dT\right) \geq \int \alpha(T) f(X(T)) dT, \quad \int \alpha(T) dT = 1, \quad \alpha(T) \geq 0 \quad \forall T.$$

Подставляя в последнее неравенство для функции  $\log$  значения  $\alpha(T) = q(T)$ ,  $X(T) = \frac{p(T)}{q(T)}$ , получаем:

$$0 = \log\left(\int q(T) \frac{p(T)}{q(T)} dT\right) \geq \int q(T) \log \frac{p(T)}{q(T)} dT = -KL(q||p).$$

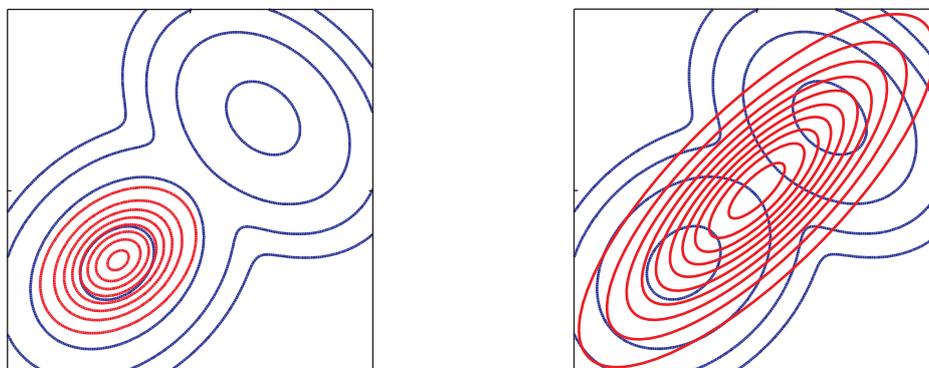
Так как коэффициенты  $q(T)$ , вообще говоря, отличны от нуля, то равенство в последнем неравенстве достигается только при  $X(T) = \text{const}$ , т.е.  $q(T) \equiv p(T)$ .

**Влияние несимметричности КЛ-дивергенции.**

Рассмотрим задачу аппроксимации распределения  $p(X)$  с помощью распределения  $q(X)$ . Будем искать приближение двумя способами:

1.  $KL(q||p) \rightarrow \min_q$ ,
2.  $KL(p||q) \rightarrow \min_q$ .

В первом случае аппроксимация ищется в области высоких значений  $q$ . Поэтому итоговое распределение, как правило, хорошо приближает распределение  $p(X)$  только на подмножестве носителя  $p(X)$  (см. рисунок ниже, случай а). Во втором случае аппроксимация ищется сразу для всего носителя распределения  $p(X)$  (см. рисунок ниже, случай б).



(а)

(б)

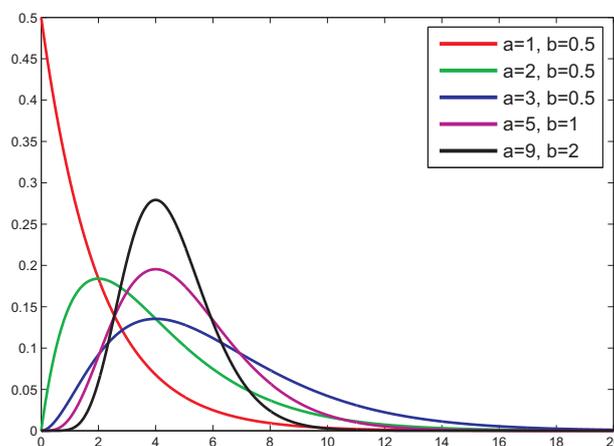
На рисунке синие кривые соответствуют линиям уровня распределения  $p(X)$ , красные кривые – распределению  $q(X)$ .

## 0.2. Гамма-распределение.

Гамма-распределение является вероятностным распределением для действительной положительной переменной  $\lambda$  и имеет плотность:

$$\mathcal{G}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda), \quad a, b > 0.$$

Здесь  $\Gamma(a)$  – гамма-функция. Различные виды гамма-распределения:



С помощью гамма-распределения можно задать широкий спектр унимодальных несимметричных распределений на положительной полуоси. Часто гамма-распределение используются в качестве априорного распределения для параметра масштаба (например, параметра  $\alpha$  в линейной и логистической регрессии). Гамма-распределение является сопряженным для параметра точности в нормальном распределении:

$$\mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right).$$

Статистики гамма-распределения:

$$\begin{aligned}\mathbb{E}\lambda &= \frac{a}{b}, \\ \mathbb{D}\lambda &= \frac{a}{b^2}, \\ \mathbb{E} \log \lambda &= \Psi(a) - \log b.\end{aligned}$$

Здесь  $\Psi(a) = \frac{d}{da} \log \Gamma(a)$  – дигамма функция.

### 1. Вывод в вероятностных моделях.

Пусть имеется некоторая вероятностная модель, задаваемая совместным распределением  $p(X, T, Z)$ . Здесь  $X$  – известные переменные,  $T$  – оцениваемые переменные,  $Z$  – неизвестные переменные, которые не требуется оценивать. Тогда задача вывода в вероятностной модели соответствует вычислению апостериорного распределения

$$p(T|X) = \frac{p(T, X)}{p(X)} = \frac{\int p(X, T, Z) dZ}{\int p(X, \tilde{T}, Z) d\tilde{T} dZ}.$$

В том случае, если нас интересуют точечные оценки, то тогда берется, как правило, мат.ожидание или мода апостериорного распределения:

$$\begin{aligned}\hat{T} &= \arg \max_T p(T|X), \\ \hat{T} &= \mathbb{E}[T|X].\end{aligned}$$

Интегралы, возникающие при получении апостериорного распределения, часто не вычисляются аналитически. Следовательно, для осуществления байесовского вывода требуются приближенные методы. Первый класс методов приближенного байесовского вывода – это методы Монте Карло с марковскими цепями. Другой класс методов – вариационный подход.

### 2. Примеры вероятностных моделей.

#### Линейная регрессия.

$$\begin{aligned}p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) &= \prod_{n=1}^N p(t_n | \mathbf{w}, \beta, \mathbf{x}_n) p(\mathbf{w} | \alpha) p(\alpha) p(\beta), \\ p(t_n | \mathbf{w}, \beta, \mathbf{x}_n) &= \mathcal{N}\left(t_n \mid \sum_{j=1}^M w_j \phi_j(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w} | \alpha) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} I), \\ p(\alpha) &= \mathcal{G}(\alpha | a_0, b_0), \\ p(\beta) &= \mathcal{G}(\beta | c_0, d_0).\end{aligned}$$

Задача вывода ( $X - (\mathbf{t}, X, \mathbf{x}_{new}), Z - (\mathbf{w}, \beta, \alpha), T - t_{new}$ ):

$$p(t_{new} | \mathbf{x}_{new}, \mathbf{t}, X) = \int p(t_{new} | \mathbf{w}, \beta, \mathbf{x}_{new}) p(\mathbf{w}, \beta, \alpha | \mathbf{t}, X) d\mathbf{w} d\beta d\alpha.$$

#### Логистическая регрессия.

$$\begin{aligned}p(\mathbf{t}, \mathbf{w}, \alpha | X) &= \prod_{n=1}^N p(t_n | \mathbf{w}, \mathbf{x}_n) p(\mathbf{w} | \alpha) p(\alpha), \\ p(t_n | \mathbf{w}, \mathbf{x}_n) &= \frac{1}{1 + \exp(-t_n \sum_{j=1}^M w_j \phi_j(\mathbf{x}_n))}, \\ p(\mathbf{w} | \alpha) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} I), \\ p(\alpha) &= \mathcal{G}(\alpha | a_0, b_0).\end{aligned}$$

Задача вывода ( $X - (\mathbf{t}, X, \mathbf{x}_{new}), Z - (\mathbf{w}, \alpha), T - t_{new}$ ):

$$p(t_{new} | \mathbf{x}_{new}, \mathbf{t}, X) = \int p(t_{new} | \mathbf{w}, \mathbf{x}_{new}) p(\mathbf{w}, \alpha | \mathbf{t}, X) d\mathbf{w} d\alpha.$$

Смесь нормальных распределений.

$$p(X, T, \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^N p(\mathbf{x}_n | \mathbf{t}_n, \boldsymbol{\mu}, \Sigma) p(\mathbf{t}_n | \boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma),$$

$$p(\mathbf{x}_n | \mathbf{t}_n, \boldsymbol{\mu}, \Sigma) = \prod_{k=1}^K [\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)]^{t_{nk}},$$

$$p(\mathbf{t}_n | \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{t_{nk}},$$

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \text{const.}$$

Задача вывода ( $X - X, Z - T, T - (\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)$ ):

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma | X) \propto p(X | \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) \rightarrow \max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma}.$$

### 3. Нижняя оценка обоснованности.

Пусть имеется вероятностная модель  $p(X, T)$  и некоторое произвольное распределение  $q(T)$ . Тогда:

$$\log p(X) = \underbrace{\int \log \frac{p(X, T)}{q(T)} q(T) dT}_{\mathcal{L}(q)} - \underbrace{\int \log \frac{p(T|X)}{q(T)} q(T) dT}_{\text{KL}(q||p(T|X))}.$$

На это равенство можно смотреть с нескольких точек зрения. Допустим, что апостериорное распределение  $p(T|X)$  не поддается вычислению (не вычисляется нормировочная константа) и мы хотим найти приближение  $q(T)$  для распределения  $p(T|X)$ . Будем искать это приближение путем минимизации КЛ-дивергенции между распределением  $q(T)$  и  $p(T|X)$  в некотором семействе распределений  $q(T)$ . Тогда из равенства выше следует, что

$$\text{KL}(q||p(T|X)) \rightarrow \min_q \Leftrightarrow \mathcal{L}(q) \rightarrow \max_q.$$

Теперь задача минимизации, которая зависит от недоступного для вычисления распределения  $p(T|X)$ , сведена к задаче максимизации функционала  $\mathcal{L}(q)$ , который зависит от известного полного совместного распределения модели  $p(X, T)$ .

С другой точки зрения, нас может интересовать значение маргинального распределения  $p(X)$  (нормировочной константы для распределения  $p(T|X)$ ). Так как  $\text{KL}(q||p(T|X)) \geq 0$ , то значение функционала  $\mathcal{L}(q)$  является нижней границей для  $\log p(X)$ . Таким образом, решая задачу максимизации функционала  $\mathcal{L}(q)$  в некотором семействе распределений  $q(T)$ , мы одновременно получаем аналитическое приближение апостериорного распределения  $p(T|X)$  и нижнюю границу для обоснованности  $\log p(X)$ .

Рассмотренные выше рассуждения справедливы также и для случая, когда требуется оценить произвольное распределение, известное с точностью до константы. Пусть имеется некоторое распределение

$$p(T) = \frac{1}{Z} \tilde{p}(T),$$

в котором мы умеем вычислять  $\tilde{p}(T)$  для произвольного  $T$ , но нормировочная константа  $Z$  является недоступной. Тогда максимизация функционала

$$\mathcal{L}(q) = \int \log \frac{\tilde{p}(T)}{q(T)} q(T) dT \rightarrow \max_q$$

позволяет найти приближение  $q(T)$  для распределения  $p(T)$ , а также оценить значение нормировочной константы  $\log Z \geq \mathcal{L}(q)$ .

### 4. Минимизация в семействе факторизованных распределений.

Рассмотрим задачу приближения апостериорного распределения  $p(T|X)$  в семействе т.н. факторизованных распределений:

$$q(T) = \prod_{j=1}^J q_j(T_j).$$

Здесь множество переменных  $T$  разбито на непересекающиеся подмножества  $T_j$ , причем  $\cup_j T_j = T$ , а  $q_j(T_j)$  – произвольное распределение в пространстве переменных  $T_j$ . Таким образом, мы приходим к следующей задаче оптимизации:

$$\mathcal{L}(q) = \int \log \frac{p(X, T)}{\prod_j q_j(T_j)} \prod_j q_j(T_j) dT_j \rightarrow \max_{q_1, \dots, q_J} .$$

Рассмотрим решение этой задачи оптимизации с помощью покоординатного подъема, т.е. зафиксируем все компоненты распределения  $q$ , кроме  $q_i$ , и рассмотрим оптимизацию  $\mathcal{L}(q)$  по отдельной компоненте  $q_i(T_i)$ . Оказывается, что решение такой (вариационной) задачи оптимизации можно получить аналитически:

$$\begin{aligned} \mathcal{L}(q) &= \int \log p(X, T) \prod_j q_j(T_j) dT_j - \sum_{j=1}^J \int \log q_j(T_j) q_j(T_j) dT_j = \\ &= \int \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right) q_i(T_i) dT_i - \int \log q_i(T_i) q_i(T_i) dT_i + \text{const}. \end{aligned}$$

Рассмотрим следующее распределение  $r(T_i)$ :

$$r(T_i) = \frac{1}{Z} \exp \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right),$$

где  $Z$  – нормировочная константа распределения, не зависящая от  $q_i(T_i)$ . Тогда  $\log r(T_i) + \log Z = \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j$ . Подставляя этот результат в выражение для  $\mathcal{L}(q)$ , получаем:

$$\begin{aligned} \mathcal{L}(q) &= \int \log r(T_i) q_i(T_i) dT_i - \int \log q_i(T_i) q_i(T_i) dT_i + \text{const} = \\ &= \int \log \frac{r(T_i)}{q_i(T_i)} q_i(T_i) dT_i + \text{const} = -\text{KL}(q_i || r) + \text{const}. \end{aligned}$$

Таким образом, максимизация  $\mathcal{L}(q)$  в данном случае эквивалентна минимизации  $\text{KL}(q_i || r)$  по распределению  $q_i(T_i)$ . Однако, известно, что минимум КЛ-дивергенции достигается в случае тождественных распределений, т.е.

$$q_i(T_i) = \frac{\exp \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right)}{\int \exp \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right) dT_i}. \quad (1)$$

Последняя формула является основным результатом вариационного подхода. Заметим, что оптимальное распределение  $q_i(T_i)$  зависит от всех остальных распределений  $q_j(T_j)$  для  $j \neq i$ . Поэтому при применении вариационного подхода возникает итерационная схема, в которой последовательно пересчитываются отдельные компоненты факторизованного распределения  $q(T)$ . При этом на каждом шаге итерации происходит монотонное увеличение нижней границы  $\mathcal{L}(q)$  для обоснованности  $\log p(X)$ , и итерационная оптимизация происходит до сходимости по значению  $\mathcal{L}(q)$ . Заметим также, что в построениях выше не накладывалось никаких ограничений на семейство распределений  $q(T)$ , кроме факторизации.

В вариационном подходе требуется уметь усреднять логарифм совместного распределения  $\log p(X, T)$  по всем компонентам  $q(T)$ , кроме одного,  $\exp \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right)$ , вычислять нормировочную константу для очередного распределения  $\int \exp \left( \int \log p(X, T) \prod_{j \neq i} q_j(T_j) dT_j \right) dT_i$ , а также нижнюю границу  $\mathcal{L}(q) = \int \log p(X, T) q(T) dT - \int \log q(T) q(T) dT$ . Заметим, что все эти величины требуют интегрирования по пространству переменных  $T$ . Здесь может создаться впечатление, что вариационный подход никак не облегчает исходную задачу поиска  $p(X)$ , т.к.  $p(X)$  также представляет собой схожий интеграл по пространству  $T \int p(X, T) dT$ . Тем не менее, во многих реальных вероятностных моделях вариационный подход действительно позволяет решить

поставленную задачу. Это связано с тем, что в вариационном подходе происходит интегрирование логарифма совместного распределения  $\log p(X, T)$ , а не самого исходного распределения  $p(X, T)$ . Кроме того, интегрирование также облегчает предположение о факторизации  $q(T)$ , т.е. интеграл часто удается разбить на произведение интегралов от подмножеств переменных  $T$ . Конкретные примеры применения вариационного подхода будут рассмотрены ниже.

Заметим, что в общей задаче байесовского вывода (см. пункт 2) мы имеем дело с тремя группами переменных  $X, T, Z$ , где  $X$  – наблюдаемые переменные,  $T$  – переменные, подлежащие оцениванию, и  $Z$  – остальные ненаблюдаемые переменные. При применении вариационного подхода происходит поиск приближения апостериорного распределения  $p(T, Z|X)$  в семействе факторизованных распределений  $q(T, Z) = q_T(T)q_Z(Z)$ . Затем

$$p(T|X) = \int p(T, Z|X)dZ \simeq \int q_T(T)q_Z(Z)dTdZ = q_T(T).$$

### 5. Пример применения вариационного подхода для модели линейной регрессии.

Рассмотрим задачу регрессии. Пусть имеется некоторая выборка  $(\mathbf{t}, X) = \{t_n, \mathbf{x}_n\}_{n=1}^N$ , состоящая из  $N$  объектов, где каждый объект представлен своим вектором признаков  $\mathbf{x}_n \in \mathbb{R}^d$  и значением регрессионной переменной  $t_n \in \mathbb{R}$ . Задача состоит в прогнозе регрессионной компоненты  $t_{new}$  для объекта, представленного только своим вектором признаков  $\mathbf{x}_{new}$ .

Для решения этой задачи воспользуемся байесовской моделью линейной регрессии:

$$p(\mathbf{t}, t_{new}, \mathbf{w}, \alpha, \beta | X, \mathbf{x}_{new}) = \prod_{n=1}^N p(t_n | \mathbf{w}, \beta, \mathbf{x}_n) p(t_{new} | \mathbf{w}, \beta, \mathbf{x}_{new}) p(\mathbf{w} | \alpha) p(\alpha) p(\beta),$$

$$p(t_n | \mathbf{w}, \sigma, \mathbf{x}_n) = \mathcal{N}\left(t_n \mid \sum_{j=1}^M w_j \phi_j(\mathbf{x}_n), \beta^{-1}\right),$$

$$p(\mathbf{w} | \alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} I),$$

$$p(\alpha) = \mathcal{G}(\alpha | a_0, b_0),$$

$$p(\beta) = \mathcal{G}(\beta | c_0, d_0).$$

С помощью вариационного подхода будет искать приближение для апостериорного распределения  $p(\mathbf{w}, \alpha, \beta | \mathbf{t}, X)$  в семействе факторизованных распределений  $q_{\mathbf{w}}(\mathbf{w})q_{\alpha}(\alpha)q_{\beta}(\beta)$ . Для этого воспользуемся общим результатом (1) и рассмотрим применение этой формулы для каждой компоненты распределения  $q_{\mathbf{w}}(\mathbf{w})$ ,  $q_{\alpha}(\alpha)$  и  $q_{\beta}(\beta)$ .

**Компонента  $q_{\mathbf{w}}(\mathbf{w})$ .**

$$\begin{aligned} \log q_{\mathbf{w}}(\mathbf{w}) &= \int \log p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) q_{\alpha}(\alpha) q_{\beta}(\beta) d\alpha d\beta + \text{const} = \\ &= \underbrace{\frac{N}{2}(\log \beta - \log 2\pi) - \frac{\mathbb{E}\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2}_{\text{не зависит от } \mathbf{w}} + \underbrace{\frac{M}{2}(\log \alpha - \log 2\pi) - \frac{\mathbb{E}\alpha}{2} \mathbf{w}^T \mathbf{w}}_{\text{не зависит от } \mathbf{w}} + \\ &+ \underbrace{\log \mathcal{G}(\alpha | a_0, b_0) + \log \mathcal{G}(\beta | c_0, d_0)}_{\text{не зависит от } \mathbf{w}} + \text{const} = -\frac{\mathbb{E}\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 - \frac{\mathbb{E}\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{const}. \quad (2) \end{aligned}$$

Здесь под  $\mathbb{E}\beta$  понимается мат. ожидание по распределению  $q_{\beta}(\beta)$ , а под  $\mathbb{E}\alpha$  – мат.ожидание по распределению  $q_{\alpha}(\alpha)$ . Выражение (2) как функция от  $\mathbf{w}$  представляет собой квадратичную функцию. Следовательно, распределение  $q_{\mathbf{w}}(\mathbf{w})$  является нормальным распределением со следующими параметрами:

$$\Sigma = \left( \text{diag}(\mathbb{E}\alpha) + \mathbb{E}\beta \sum_n \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right)^{-1}, \quad (3)$$

$$\boldsymbol{\mu} = \Sigma \left( \sum_n \mathbb{E}\beta t_n \phi(\mathbf{x}_n) \right). \quad (4)$$

Компонента  $q_\alpha(\alpha)$ .

$$\begin{aligned} \log q_\alpha(\alpha) &= \int \log p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) q_{\mathbf{w}}(\mathbf{w}) q_\beta(\beta) d\mathbf{w} d\beta + \text{const} = \\ &= \frac{M}{2} \log \alpha + \frac{\alpha}{2} \mathbb{E} \mathbf{w}^T \mathbf{w} + (a_0 - 1) \log \alpha - b_0 \alpha + \text{const}. \end{aligned}$$

В этом выражении легко узнается логарифм гамма-распределения с параметрами:

$$a = a_0 + \frac{M}{2}, \quad (5)$$

$$b = b_0 + \mathbb{E} \mathbf{w}^T \mathbf{w}. \quad (6)$$

Компонента  $q_\beta(\beta)$ .

$$\begin{aligned} \log q_\beta(\beta) &= \int \log p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) q_{\mathbf{w}}(\mathbf{w}) q_\alpha(\alpha) d\mathbf{w} d\alpha + \text{const} = \\ &= \frac{N}{2} \log \beta - \frac{\beta}{2} \sum_{n=1}^N \mathbb{E} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + (c_0 - 1) \log \beta - d_0 \beta + \text{const}. \end{aligned}$$

В этом выражении также узнается логарифм гамма-распределения с параметрами:

$$c = c_0 + \frac{N}{2}, \quad (7)$$

$$d = d_0 + \frac{1}{2} \sum_n \mathbb{E} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2. \quad (8)$$

**Итоговый алгоритм.**

Таким образом, на каждой итерации вариационного алгоритма компоненты факторизованного распределения  $q(\mathbf{w}, \alpha, \beta)$  представляют собой

$$q_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \Sigma), \quad (9)$$

$$q_\alpha(\alpha) = \mathcal{G}(\alpha | a, b), \quad (10)$$

$$q_\beta(\beta) = \mathcal{G}(\beta | c, d), \quad (11)$$

где соответствующие параметры распределений пересчитываются по формулам (3)-(4), (5)-(6) и (7)-(8). При этом необходимые статистики распределений для формул пересчета выглядят следующим образом:

$$\mathbb{E} \alpha = \frac{a}{b},$$

$$\mathbb{E} \beta = \frac{c}{d},$$

$$\mathbb{E} \mathbf{w}^T \mathbf{w} = \text{tr} \mathbb{E}(\mathbf{w} \mathbf{w}^T),$$

$$\mathbb{E} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 = t_n^2 - 2 \mathbb{E} \mathbf{w}^T \phi(\mathbf{x}_n) + \phi(\mathbf{x}_n)^T \mathbb{E} \mathbf{w} \mathbf{w}^T \phi(\mathbf{x}_n),$$

$$\mathbb{E} \mathbf{w} = \boldsymbol{\mu},$$

$$\mathbb{E} \mathbf{w} \mathbf{w}^T = \Sigma + \boldsymbol{\mu} \boldsymbol{\mu}^T.$$

Заметим, что по аналогии со схемой Гиббса в вариационном подходе все компоненты распределения пересчитываются последовательно. Сначала фиксируются начальные приближения для всех параметров распределений  $\boldsymbol{\mu}, \Sigma, a, b, c, d$ . Затем, новые значения  $\boldsymbol{\mu}$  и  $\Sigma$  вычисляются по формулам (3)-(4). Эти значения используются для вычисления необходимых статистик распределения  $\mathbb{E} \mathbf{w}$ ,  $\mathbb{E} \mathbf{w} \mathbf{w}^T$  и др. После этого находятся новые значения  $a$  и  $b$  по формулам (5)-(6) и новые статистики  $\mathbb{E} \alpha$  и  $\mathbb{E} \log \alpha$ . И так далее.

Как уже было отмечено выше, в вариационном подходе помимо приближения апостериорного распределения  $p(\mathbf{w}, \alpha, \beta | \mathbf{t}, X) \simeq q_{\mathbf{w}}(\mathbf{w})q_{\alpha}(\alpha)q_{\beta}(\beta)$  мы можем получить также нижнюю границу на обоснованность  $\log p(\mathbf{t} | X)$ :

$$\log p(\mathbf{t} | X) \geq \mathcal{L}(q) = \mathbb{E} \log p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) - \mathbb{E} \log q_{\mathbf{w}}(\mathbf{w}) - \mathbb{E} \log q_{\alpha}(\alpha) - \mathbb{E} \log q_{\beta}(\beta).$$

С учетом результата (9)-(11) нижняя граница  $\mathcal{L}(q)$  вычисляется аналитически:

$$\begin{aligned} \mathbb{E} \log p(\mathbf{t}, \mathbf{w}, \alpha, \beta | X) &= \frac{N}{2} (\mathbb{E} \log \beta - \log 2\pi) - \frac{\mathbb{E} \beta}{2} \sum_n \mathbb{E} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \frac{M}{2} (\mathbb{E} \log \alpha - \log 2\pi) - \frac{\mathbb{E} \alpha}{2} \mathbb{E} \mathbf{w}^T \mathbf{w} + \\ &+ (a_0 - 1) \mathbb{E} \log \alpha - b_0 \mathbb{E} \alpha + a_0 \log b_0 - \log \Gamma(a_0) + (c_0 - 1) \mathbb{E} \log \beta - d_0 \mathbb{E} \beta + c_0 \log d_0 - \log \Gamma(c_0). \end{aligned}$$

$$\mathbb{E} \log q_{\mathbf{w}}(\mathbf{w}) = -\frac{M}{2} (\log 2\pi + 1) - \frac{1}{2} \log \det \Sigma,$$

$$\mathbb{E} \log q_{\alpha}(\alpha) = (a - 1) \mathbb{E} \log \alpha - b \mathbb{E} \alpha + a \log b - \log \Gamma(a),$$

$$\mathbb{E} \log q_{\beta}(\beta) = (c - 1) \mathbb{E} \log \beta - d \mathbb{E} \beta + c \log d - \log \Gamma(c).$$

Для вычисления этих выражений необходимо знать ряд дополнительных статистик распределений, которые вычисляются следующим образом (см. ликбез для гамма-распределения):

$$\mathbb{E} \log \alpha = \Psi(a) - \log(b),$$

$$\mathbb{E} \log \beta = \Psi(c) - \log(d).$$

В ходе итерационного процесса значение  $\mathcal{L}(q)$  не убывает. Итерационный процесс заканчивается, когда значение  $\mathcal{L}(q)$  стабилизируется.

Для получения прогноза регрессионной компоненты  $t_{new}$  для объекта  $\mathbf{x}_{new}$  необходимо вычислить следующий интеграл:

$$p(t_{new} | \mathbf{x}_{new}, \mathbf{t}, X) = \int p(t_{new} | \mathbf{x}_{new}, \mathbf{w}, \beta) p(\mathbf{w}, \alpha, \beta | \mathbf{t}, X) d\mathbf{w} d\alpha d\beta.$$

Подставляя вместо апостериорного распределения  $p(\mathbf{w}, \alpha, \beta | \mathbf{t}, X)$  его факторизованное приближение, получаем:

$$\begin{aligned} p(t_{new} | \mathbf{x}_{new}, \mathbf{t}, X) &\simeq \hat{p}(t_{new} | \mathbf{x}_{new}, \mathbf{t}, X) = \int p(t_{new} | \mathbf{x}_{new}, \mathbf{w}, \beta) q_{\mathbf{w}}(\mathbf{w}) q_{\beta}(\beta) d\mathbf{w} d\beta = \\ &= \int \left[ \int \mathcal{N}(t_{new} | \mathbf{w}^T \phi(\mathbf{x}_{new}), \beta^{-1}) \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \Sigma) d\mathbf{w} \right] \mathcal{G}(\beta | c, d) d\beta = \\ &= \int \mathcal{N}(t_{new} | \boldsymbol{\mu}^T \phi(\mathbf{x}_{new}), \beta^{-1} + \phi(\mathbf{x}_{new})^T \Sigma \phi(\mathbf{x}_{new})) \mathcal{G}(\beta | c, d) d\beta. \end{aligned}$$

Последний интеграл не берется аналитически. Однако, данный интеграл является одномерным и поэтому может быть эффективно оценен с помощью метода Монте Карло, т.к. мы можем легко получать выборку из гамма-распределения. Кроме того, можно показать, что  $\mathbb{E}_{\hat{p}} t_{new} = \boldsymbol{\mu}^T \phi(\mathbf{x}_{new})$ , а  $\mathbb{D}_{\hat{p}} t_{new} \simeq (\mathbb{E} \beta)^{-1} + \phi(\mathbf{x}_{new})^T \Sigma \phi(\mathbf{x}_{new})$ .