ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ II

Сергей Николенко

СПбГУ — Санкт-Петербург

25 апреля 2020 г.

Random facts:

- 25 апреля — День ДНК; именно 25 апреля 1953 г. в *Nature* вышли три статьи Джеймса Уотсона, Фрэнсиса Крика, Мориса Уилкинса, Розалинд Франклин и их коллег, а 25 апреля 2003 г. объявили о завершении проекта «Геном человека»
- 25 апреля 1792 г. разбойник Николя Пеллетье первым испробовал на себе гильотину
- 25 апреля 1945 г. советские и американские войска встретились на Эльбе
- 25 апреля 1960 г. субмарина *USS Triton* завершила первое подводное кругосветное плавание
- 25 апреля 1983 г. Юрий Андропов пригласил в Советский Союз Саманту Смит
- 25 апреля 1993 г. на референдуме в России большинство поддержали политику Бориса Ельцина, а 25 апреля 2007 г. прошли его похороны

Вариационное приближение для гауссиана

Одномерный гауссиан

- И ещё пример: давайте найдём параметры одномерного гауссиана по точкам $\mathbf{X} = \{x_1, \dots, x_N\}$. Правдоподобие:

$$p(\mathbf{X} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} e^{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2}.$$

- Вводим сопряжённые априорные распределения:

$$p(\mu \mid \tau) = N(\mu \mid \mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = \mathrm{Gamma}(\tau \mid a_0, b_0).$$

- Мы это только что подсчитали точно, но давайте приблизим теперь апостериорное распределение как

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau).$$

- На самом деле так не раскладывается!
- Это то, что мы делали для $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$. Посчитаем...

Одномерный гауссиан

- ...$q_\mu(\mu)$ – гауссиан с параметрами

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N)\mathrm{E}[\tau].$$

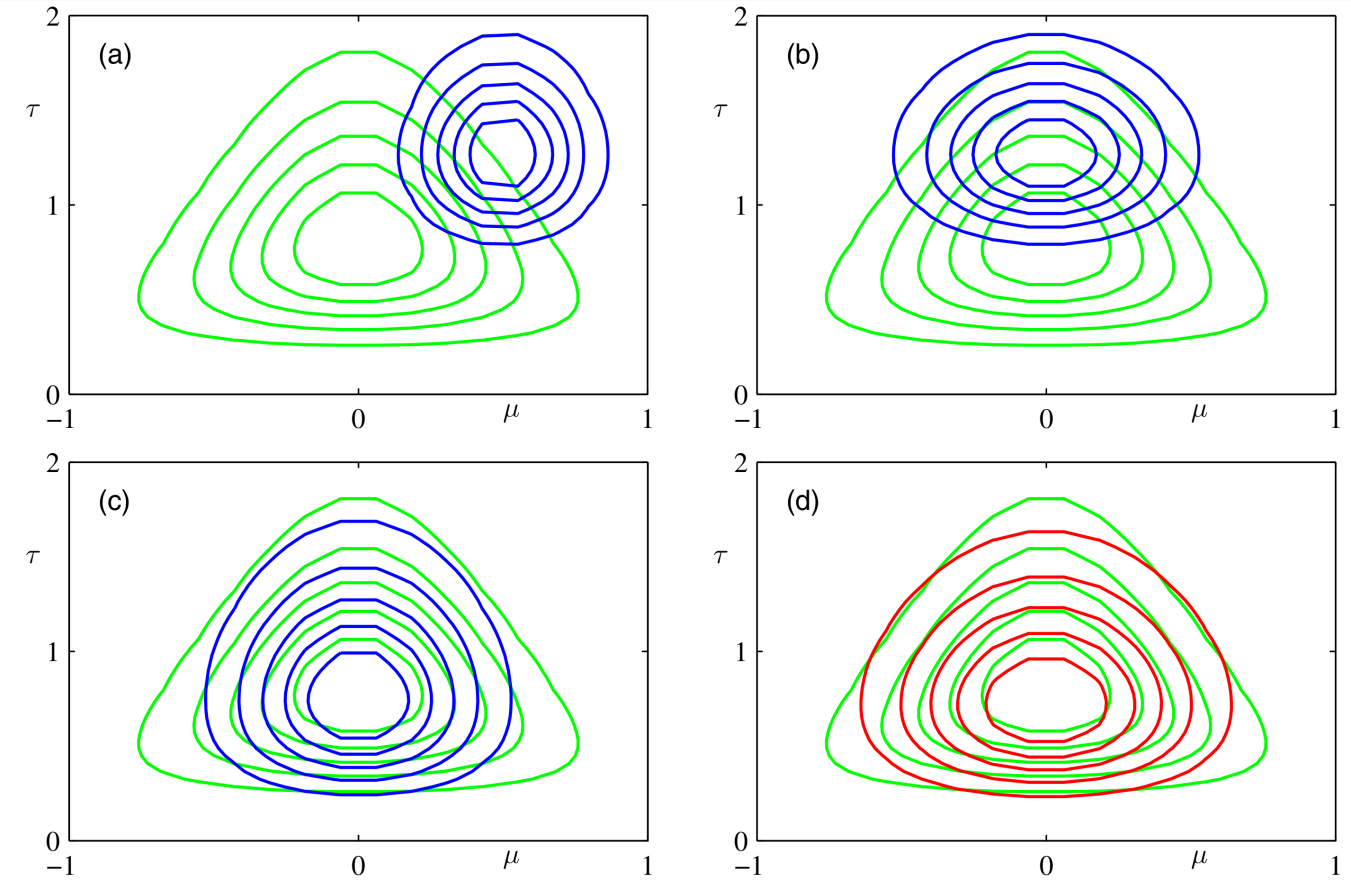
- А $q_\tau(\tau)$ – гамма-распределение с параметрами

$$a_N = a_0 + \frac{N+1}{2}, \quad b_N = b_0 + \frac{1}{2}\mathrm{E}_\mu\left[\sum_n (x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

- Всё получилось как надо, но без предположений о форме q_τ и q_μ.

Одномерный гауссиан

- Вот такой вывод в пространстве (μ, τ):



- А для $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ (non-informative priors) можно и точно посчитать...

Одномерный гауссиан

- Получатся моменты для μ

$$\mathrm{E}[\mu] = \bar{x}, \quad \mathrm{E}[\mu^2] = \bar{x}^2 + \frac{1}{N\mathrm{E}[\tau]}.$$

- Это можно подставить и найти $\mathrm{E}[\tau]$:

$$\frac{1}{\mathrm{E}[\tau]} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

Вариационное приближение для смеси гауссианов

Смесь гауссианов

- Смесь гауссианов: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$,

$$p(\mathbf{Z} \mid \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} N\left(\mathbf{x}_n \mid \mu_k, \Lambda_k^{-1}\right).$$

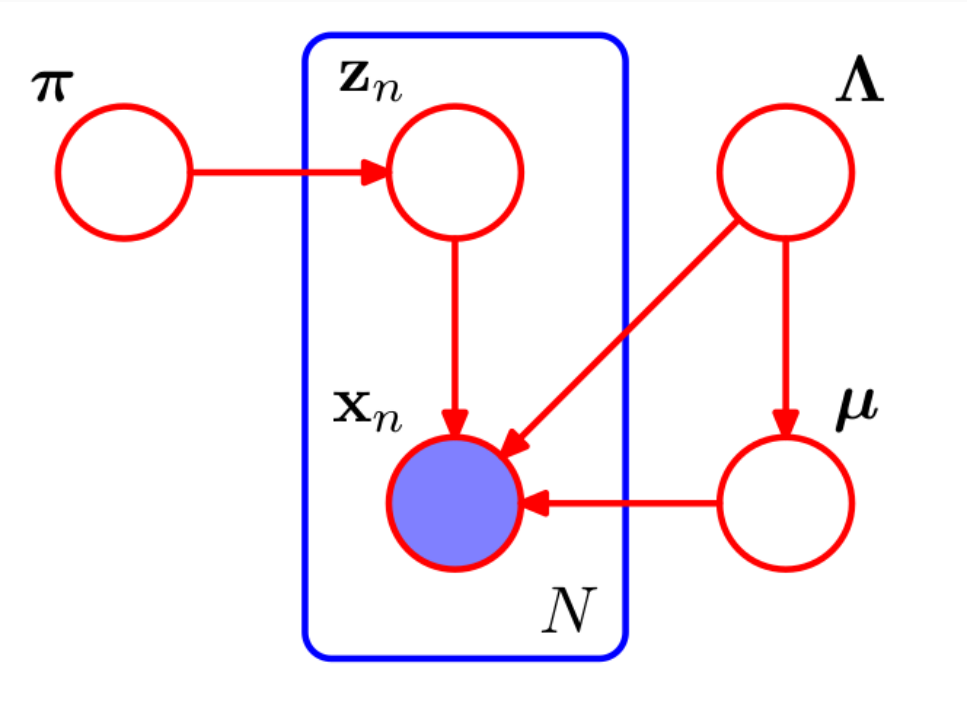
- Выберем сопряжённые априорные распределения:

$$p(\pi) = \mathrm{Dir}(\pi \mid \alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1},$$

$$\begin{aligned} p(\mu, \Lambda) &= p(\mu \mid \Lambda) p(\Lambda) \\ &= \prod_{k=1}^{K} N\left(\mu_k \mid \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}\right) W(\Lambda_k \mid W_0, \nu_0). \end{aligned}$$

Смесь гауссианов

- Вот такая графическая модель:



- Распределение Дирихле пусть будет симметричное для простоты; часто ещё $\mathbf{m}_0 = 0$.
- Заметьте разницу между латентными переменными и параметрами модели.

Вариационное приближение

- Теперь вариационное приближение. Сначала сама факторизация:

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda)p(\mathbf{Z} \mid \pi)p(\pi)p(\mu \mid \Lambda)p(\Lambda).$$

- Мы наблюдаем только $\mathbf{X}$, остальное всё надо как-то оценить.
- Интересно, что единственное предположение про наше вариационное приближение выглядит так:

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda).$$

- И всё! Дальше всё само собой получится. Но не сразу...

Вариационное приближение

- Сначала $q^*(\mathbf{Z})$:

$$\begin{aligned}\ln q^*(\mathbf{Z}) &= \mathrm{E}_{\pi,\mu,\Lambda}\left[\ln p(\mathbf{X},\mathbf{Z},\pi,\mu,\Lambda)\right] + \text{const} \\ &= \mathrm{E}_{\pi,\mu,\Lambda}\left[\ln p(\mathbf{Z}\mid\pi)\right] + \mathrm{E}_{\mu,\Lambda}\left[\ln p(\mathbf{X}\mid\mathbf{Z},\mu,\Lambda)\right] + \text{const} \\ &= \ldots = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + \text{const},\end{aligned}$$

$$\text{где } \ln\rho_{nk} = \mathrm{E}[\ln\pi_k] + \frac{1}{2}\mathrm{E}[\ln|\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \\ - \frac{1}{2}\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{x}_n-\mu_k)^\top\Lambda_k(\mathbf{x}_n-\mu_k)\right].$$

- Нормируем:

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}, \text{ где } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}.$$

Вариационное приближение

- Теперь $\mathrm{E}[z_{nk}] = r_{nk}$, т.е. r_{nk} – то, насколько точка $\mathbf{x}_n$ принадлежит кластеру k.
- Можно определить статистики с их учётом, как обычно:

$$N_k = \sum_{n=1}^{N} r_{nk},$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n,$$

$$S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top.$$

- То же самое происходило и в EM-алгоритме.

Вариационное приближение

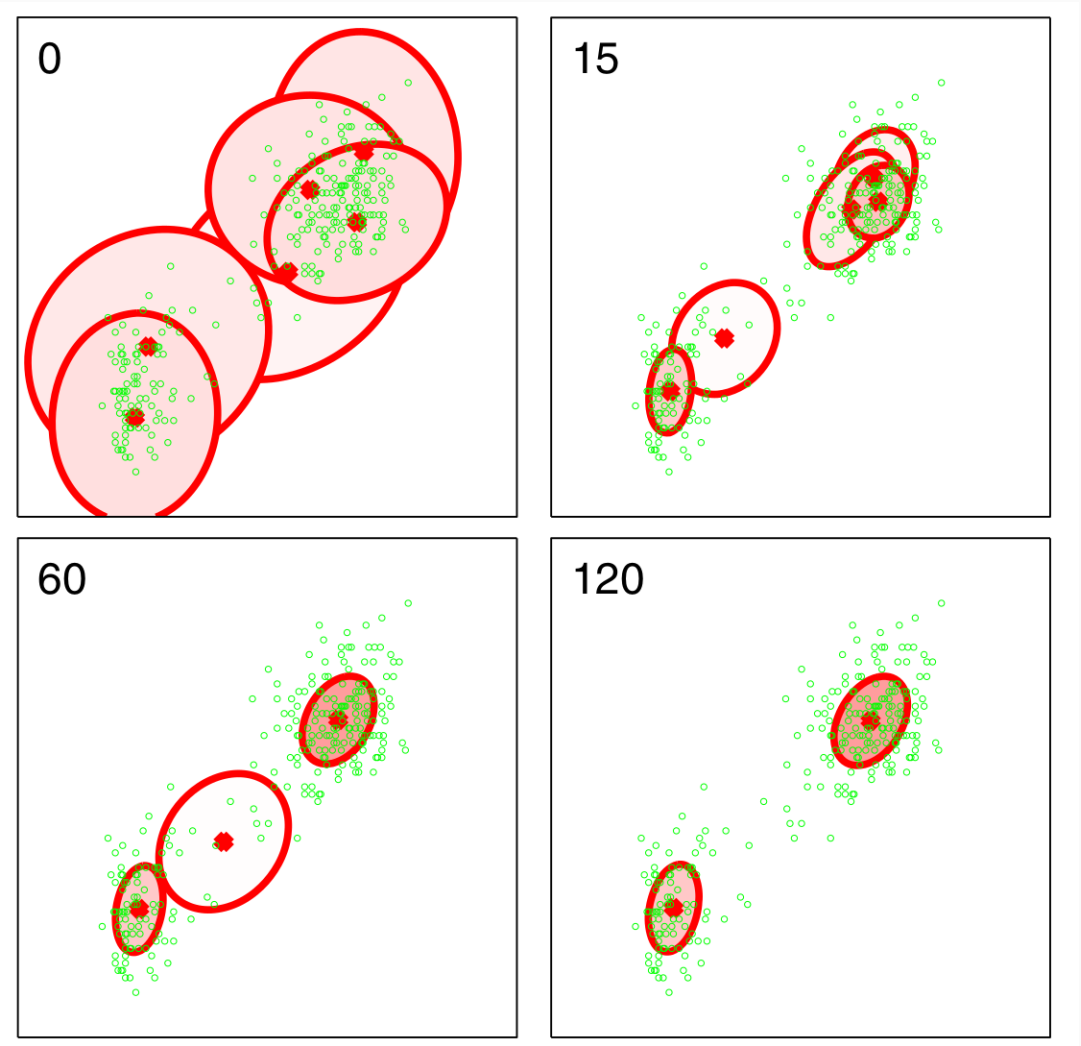
- Теперь $q^*(\pi, \mu, \Lambda)$:

$$\ln q^*(\pi, \mu, \Lambda) = \ln p(\pi) + \sum_{k=1}^{K} \ln p(\mu_k, \Lambda_k) + \mathrm{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \pi)]$$
$$+ \sum_{k=1}^{K} \sum_{n=1}^{N} \mathrm{E}[z_{nk}] \ln N(\mathbf{x}_n \mid \mu_k \Lambda_k^{-1}) + \mathrm{const}.$$

- Вот уже получилось, что $q^*(\pi, \mu, \Lambda)$ раскладывается в $q^*(\pi)q^*(\mu, \Lambda)$, опять же без предположений.
- Более того, $q^*(\mu, \Lambda) = \prod_{k=1}^{K} q(\mu_k, \Lambda_k)$.
- И теперь можно по отдельности посчитать (упражнение), получится типичный M-шаг.
- Причём распределения останутся той же формы (т.к. были сопряжённые).

Вариационное приближение

- Теперь даже model selection автоматически получается, просто у некоторых компонент $N_k \approx 0$:



- Никакого оверфиттинга или коллапса компонент.

Вариационное приближение

- А ещё полезно уметь считать саму вариационную нижнюю оценку

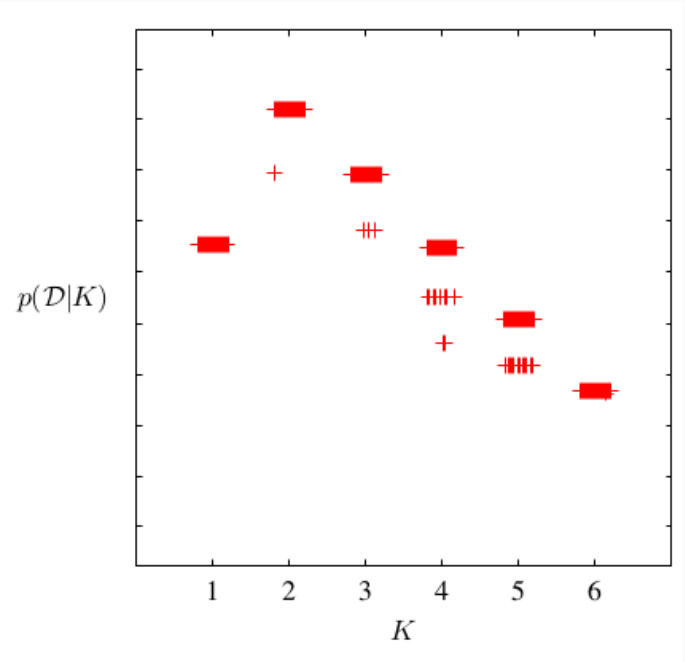
$$\mathcal{L}(q) = \int q(Z) \ln \frac{p(X, Z)}{q(Z)} \mathrm{d}Z = \mathbb{E}_q \left[\ln p(X, Z)\right] - \mathbb{E}_q \left[\ln q(Z)\right].$$

- В данном случае

$$\begin{aligned}\mathcal{L} = \mathbb{E}\left[\ln p(X, Z, \pi, \mu, \Lambda)\right] - \mathbb{E}\left[\ln q(Z, \pi, \mu, \Lambda)\right] = \\ = \mathbb{E}\left[\ln p\left(X \mid Z, \mu, \Lambda\right)\right] + \mathbb{E}\left[\ln p\left(Z \mid \pi\right)\right] + \mathbb{E}\left[\ln p(\pi)\right] + \mathbb{E}\left[\ln p(\mu, \Lambda)\right] - \\ - \mathbb{E}\left[\ln q(Z)\right] - \mathbb{E}\left[\ln q(\pi)\right] - \mathbb{E}\left[\ln q(\mu, \Lambda)\right].\end{aligned}$$

Вариационное приближение

- И эту оценку можно использовать тоже для выбора моделей, только надо ещё $\ln K!$ добавить.



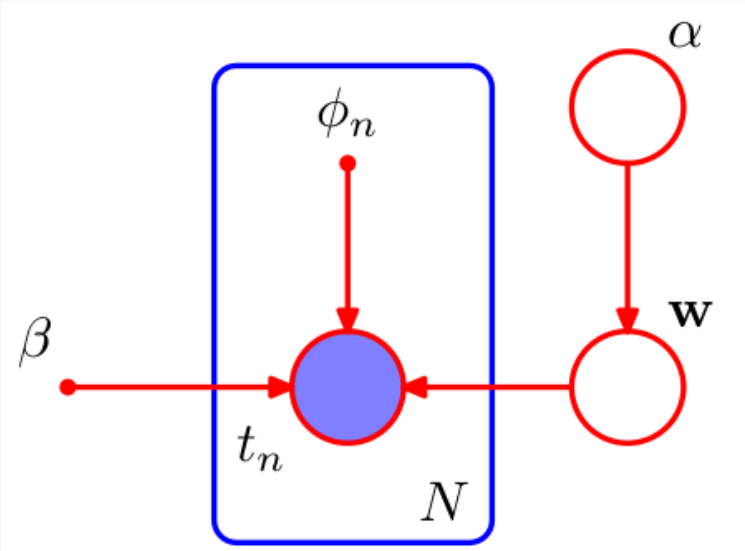
- А можно проще: оценивать и пересчитывать по отдельности $\pi_k = \frac{1}{N}\sum_{n=1}^{N} r_{nk}$, максимизируя нижнюю оценку, и для некоторых компонент получим $\pi_k \to 0$, надо просто начать с достаточно большого числа K.

Линейная регрессия

- Для линейной регрессии:

$$p\left(\mathbf{y} \mid \mathbf{w}\right) = \prod_{n=1}^{N} N(\mathbf{w}^\top \mathbf{x}_n, \beta^{-1}), \quad p\left(\mathbf{w} \mid \alpha\right) = N(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I}).$$

- Можем ввести теперь априорное распределение на α: $p(\alpha) = \mathrm{Gamma}(\alpha \mid a_0, b_0)$.



Линейная регрессия

- Хотим найти $p(\mathbf{w}, \alpha \mid \mathbf{y})$, вводим вариационное приближение:

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha), \text{ и теперь}$$

$$\begin{aligned}\ln q^*(\alpha) &= \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}\left[\ln p\left(\mathbf{w} \mid \alpha\right)\right] + \text{const} \\ &= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{d}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^\top\mathbf{w}\right] + \text{const},\end{aligned}$$

и это гамма-распределение с параметрами

$$a_N = a_0 + \frac{d}{2}, \quad b_N = b_0 + \frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^\top\mathbf{w}\right].$$

Линейная регрессия

- Аналогично,

$$\begin{aligned}\ln q^*(\mathbf{w}) &= \ln p(\mathbf{y} \mid \mathbf{w}) + \mathbb{E}_\alpha [\ln p(\mathbf{w} \mid \alpha)] + \text{const} \\ &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^\top \mathbf{x}_n)^2 - \frac{1}{2}\mathbb{E}[\alpha]\,\mathbf{w}^\top\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^\top(\mathbb{E}[\alpha]\,\mathbf{I} + \beta X^\top X)\,\mathbf{w} + \beta\mathbf{w}^\top X^\top \mathbf{y} + \text{const},\end{aligned}$$

и это, конечно, гауссиан с параметрами

$$S_N = (\mathbb{E}[\alpha]\,\mathbf{I} + \beta X^\top X)^{-1}, \quad \mathbf{m}_N = \beta S_N X^\top \mathbf{I}.$$

Линейная регрессия

- Стандартная теория вероятностей говорит, что

$$\mathbb{E}[\alpha] = a_N / b_N, \quad \mathbb{E}\left[\mathbf{w}^\top \mathbf{w}\right] = \mathbf{m}_N \mathbf{m}_N^\top + S_N.$$

- И теперь можно запускать итеративный алгоритм пересчёта. А если подставить $a_0 = b_0 = 0$, получится

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{d/2}{\mathbb{E}\left[\mathbf{w}^\top \mathbf{w}\right]/2} = \frac{d}{\mathbf{m}_N^\top \mathbf{m}_N + \mathrm{Tr}(S_N)}.$$

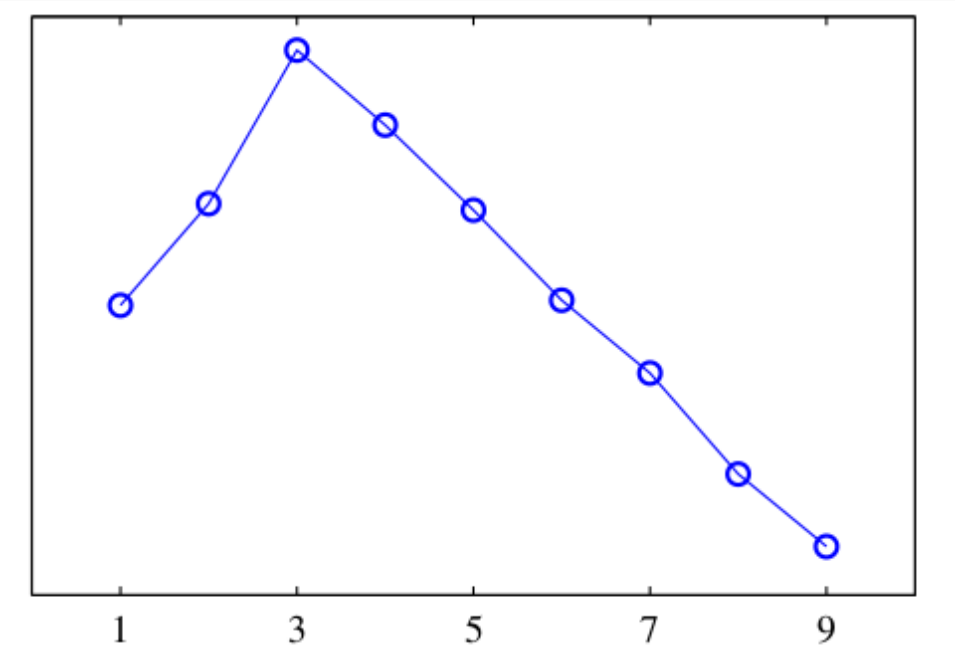
Линейная регрессия

- Можно и саму нижнюю оценку посчитать:

$$\mathcal{L}(q) = \mathbb{E}\left[\ln p(\mathbf{w}, \alpha, \mathbf{y})\right] - \mathbb{E}\left[\ln q(\mathbf{w}, \alpha)\right] =$$
$$\mathbb{E}_{\mathbf{w}}\left[\ln p\left(\mathbf{y} \mid \mathbf{w}\right)\right] + \mathbb{E}_{\mathbf{w},\alpha}\left[\ln p\left(\mathbf{w} \mid \alpha\right)\right] + \mathbb{E}_{\alpha}\left[\ln p(\alpha)\right] -$$
$$- \mathbb{E}_{\mathbf{w}}\left[\ln q(\mathbf{w})\right] - \mathbb{E}_{\alpha}\left[\ln q(\alpha)\right].$$

Линейная регрессия

- И всё это подставляется из предыдущих формул.
- Пример графика нижней оценки для полиномиальной регрессии, где точки были просэмплированы из кубического многочлена:



Другие примеры

- Есть другие примеры вариационных приближений.
- Обращение матриц; например, для линейной регрессии надо посчитать $\beta^* = C^{-1}\mathbf{b}$:

$$J(\beta) = \frac{1}{2}(\beta^* - \beta)^\top C(\beta^* - \beta) = \ldots = \text{Const} - \beta^\top \mathbf{b} + \frac{1}{2}\beta^\top C\beta,$$

  и теперь можно решать такую задачу выпуклой оптимизации.
- Метод конечных элементов – для уравнения Пуассона $-u''(x) = f(x)$, $x \in (a, b)$:

$$J(u) = \frac{1}{2}\int_a^b \left(u'(x) - u^{*\prime}(x)\right)^2 \mathrm{d}x = \ldots = \text{Const} - \int_a^b u(x)f(x)\mathrm{d}x + \frac{1}{2}\int_a^b u'(x)^2\mathrm{d}x,$$

  и если ищем в подпространстве $\tilde{u}(x) = \sum_{i=1}^k \alpha_i\phi_i(x)$, то опять

$$\tilde{J}(\alpha) = \alpha^\top \mathbf{b} + \frac{1}{2}\alpha^\top C\alpha.$$

В графических моделях

- В графических моделях – теория среднего поля (mean field theory). Пусть дано $p(\mathbf{x})$, $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_h)$, и надо найти

$$\log p(\mathbf{x}_v) = \log \sum_{\mathbf{x}_h} p(\mathbf{x}_v, \mathbf{x}_h), \quad p(\mathbf{x}_h \mid \mathbf{x}_v) = p(\mathbf{x}_h, \mathbf{x}_v)/p(\mathbf{x}_v).$$

- Опять делаем тот же трюк:

$$J(q) = \log p(\mathbf{x}_v) - \mathrm{KL}(q_{\mathbf{x}_h} \| p_{\mathbf{x}_h|\mathbf{x}_v}) = \log p(\mathbf{x}_v) - \sum_{\mathbf{x}_h} q(\mathbf{x}_h) \log \frac{q(\mathbf{x}_h)}{p(\mathbf{x}_h \mid \mathbf{x}_v)}$$

$$\ldots = H(q) + \mathbb{E}_q \left[\log p(\mathbf{x}_h, \mathbf{x}_v)\right] = H(q) + \sum_{C \in C} \sum_{\mathbf{x}_{C \cap h}} q(\mathbf{x}_{C \cap h}) \log \Psi_C(\mathbf{x}_C),$$

где $q(\mathbf{x}_{C \cap h})$ – маргинальная вероятность по скрытым переменным из клики C.

- Теория среднего поля – это когда $q(\mathbf{x}_h) = \prod_{i \in h} q_i(x_i)$.

Спасибо!

Спасибо за внимание!