Байесовское мультимоделирование: вариационный вывод-2

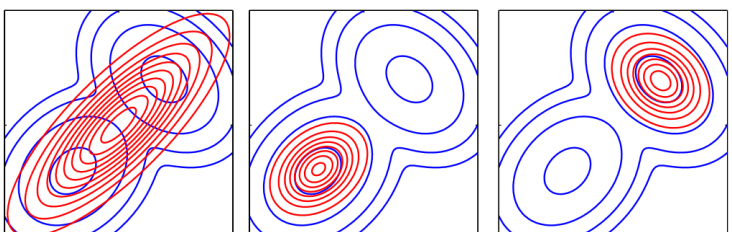
Московский Физико-Технический Институт

2021

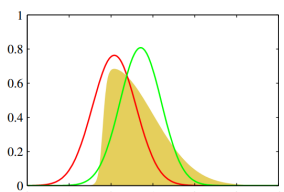
Вариационная оценка, ELBO

Вариационная оценка Evidence, Evidence lower bound — метод нахождения приближенного значения аналитически невычислимого распределения $p(\boldsymbol{w}|\mathfrak{D}, \boldsymbol{h})$ распределением $q(\boldsymbol{w}) \in \mathfrak{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\mathfrak{D})) = -\int_{\boldsymbol{w}} q(\boldsymbol{w}) \log \frac{p(\boldsymbol{w}|\mathfrak{D})}{q(\boldsymbol{w})} d\boldsymbol{w} = \mathrm{E}_{\boldsymbol{w}} \log p(\mathfrak{D}|\boldsymbol{w}) - \mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{h})).$$



Вариационный вывод и expectation propogation (Bishop)



Аппроксимация Лапласа и вариационная оценка

Вариационный вывод: наивная интерпретация

Вариационная оценка обоснованности получается путем приближения заранее определенным **параметрическим непрерывным** распределением (**нормальным**) апостериорного распределения. Задача получения вариационной оценки **сводится к минимизиации**

$$\mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\mathfrak{D})) = -\int_{\boldsymbol{w}} q(\boldsymbol{w})\log\frac{p(\boldsymbol{w}|\mathfrak{D})}{q(\boldsymbol{w})}d\boldsymbol{w} = \mathrm{E}_{\boldsymbol{w}}\log p(\mathfrak{D}|\boldsymbol{w}) - \mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{h})).$$

MCMC и вариационный вывод

Идея MCMC: Порождаем сэмплы из простого распределения и принимаем их, если заданное отношение больше порога:

$$\min\left(1, \frac{p(\boldsymbol{w}^{\tau}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})}{p(\boldsymbol{w}^{\tau-1}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})}\right),$$

где $\boldsymbol{w}^{\tau}$ выбирается на основе предыдущего сэмпла:

$$\boldsymbol{w}^{\tau} = T(\boldsymbol{w}^{\tau-1}).$$

Salimans et al., 2014: будем интерпретировать последовательность применения оператора T как оптимизацию вариационной оценки:

$$T^1 \circ \dots T^{\eta}(\boldsymbol{w}) \to p(\boldsymbol{w}^{\tau}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h}).$$

Maclaurin et. al, 2015: в качестве оператора T будем рассматривать оператор оптимизации. Откажемся от отклонения сэмплов по порогу.

Оператор оптимизации, Maclaurin et. al, 2015

Определение

Назовем оператором оптимизации алгоритм T выбора вектора параметров $\boldsymbol{w}'$ по параметрам предыдущего шага $\boldsymbol{w}$:

$$\boldsymbol{w}' = T(\boldsymbol{w}).$$

Определение

Пусть L — дифференцируемая функция потерь.
Оператором градиентного спуска назовем следующий оператор:

$$T(\boldsymbol{w}) = \boldsymbol{w} - \beta \nabla L(\boldsymbol{w}, \boldsymbol{y}, \mathfrak{D}).$$

Градиентный спуск для оценки правдоподобия

Рассмотрим максимизацию совместного распределения параметров:

$$L = -\log\, p(\mathfrak{D}, \boldsymbol{w}|\boldsymbol{h}) = -\sum_{\mathfrak{D}\in\mathfrak{D}} \log\, p(\mathfrak{D}|\boldsymbol{w}, \boldsymbol{h}) p(\boldsymbol{w}|\boldsymbol{h})$$

Проведем оптимизацию нейросети из r различных начальных приближений $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$ с использованием градиентного спуска:

$$\boldsymbol{w}' = T(\boldsymbol{w}).$$

Векторы параметров $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$ соответствуют некоторому скрытому распределению $q(\boldsymbol{w})$.

Энтропия

Формулу вариационной оценки можно переписать с использованием энтропии:

$$\log\, p(\mathfrak{D}|\boldsymbol{f}) \geq \int_{\boldsymbol{w}} q(\boldsymbol{w}) \log\, \frac{p(\mathfrak{D}, \boldsymbol{w}|\boldsymbol{h})}{q(\boldsymbol{w})} d\boldsymbol{w} =$$

$$\mathrm{E}_{q(\boldsymbol{w})}[\log\, p(\mathfrak{D}, \boldsymbol{w}|\boldsymbol{h})] + \mathrm{S}(q(\boldsymbol{w})),$$

где $\mathrm{S}(q(\boldsymbol{w}))$ — энтропия:

$$\mathrm{S}(q(\boldsymbol{w})) = -\int_{\boldsymbol{w}} q(\boldsymbol{w}) \log\, q(\boldsymbol{w}) d\boldsymbol{w}.$$

Градиентный спуск для оценки правдоподобия

Утверждение 3

Пусть L — липшицева функция, оператор оптимизации — биекция. Тогда разность энтропии на различных шагах оптимизации вычисляется как:

$$\mathrm{S}(q'(\boldsymbol{w})) - \mathrm{S}(q(\boldsymbol{w})) \simeq \frac{1}{r}\sum_{g=1}^{r}\big(-\beta\mathrm{Tr}[\boldsymbol{H}(\boldsymbol{w}'^{g})] - \beta^2\mathrm{Tr}[\boldsymbol{H}(\boldsymbol{w}'^{g})\boldsymbol{H}(\boldsymbol{w}'^{g})]\big).$$

Итоговая оценка на шаге оптимизации τ:

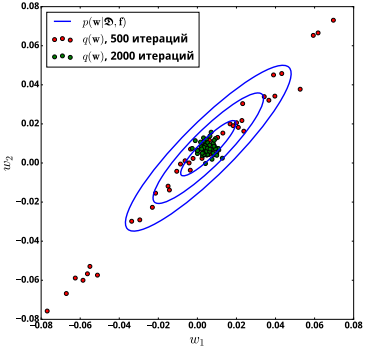
$$\log\hat{p}(\boldsymbol{Y}|\mathfrak{D},\boldsymbol{h}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\boldsymbol{w}_{\tau}^{g},\mathfrak{D},\boldsymbol{Y}) + \mathrm{S}(q^0(\boldsymbol{w}))+$$

$$+\frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\big(-\beta\mathrm{Tr}[\boldsymbol{H}(\boldsymbol{w}_{b}^{g})] - \beta^2\mathrm{Tr}[\boldsymbol{H}(\boldsymbol{w}_{b}^{g})\boldsymbol{H}(\boldsymbol{w}_{b}^{g})]\big),$$

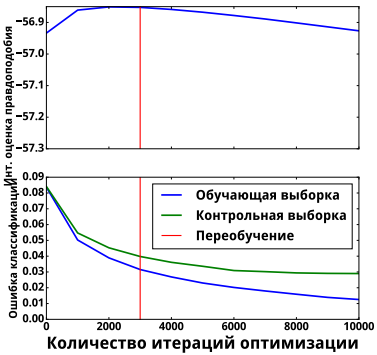
$\boldsymbol{w}_{b}^{g}$ — вектор параметров старта g на шаге b, $\mathrm{S}(q^0(\boldsymbol{w}))$ — начальная энтропия.

Переобучение, Maclaurin et. al, 2015

Градиентный спуск не минимизирует дивергенцию $\mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\mathfrak{D},\boldsymbol{h}))$. При приближении к моде распределения снижается оценка Evidence, что интерпретируется как переоубчение модели.



Схождение распределения к моде



Оценка начала переобучения

Стохастическая динамика Ланжевена

Модификация стохастического градиентного спуска:

$$T = \boldsymbol{w} - \beta \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

где шаг оптимизации α изменяется с количеством итераций:

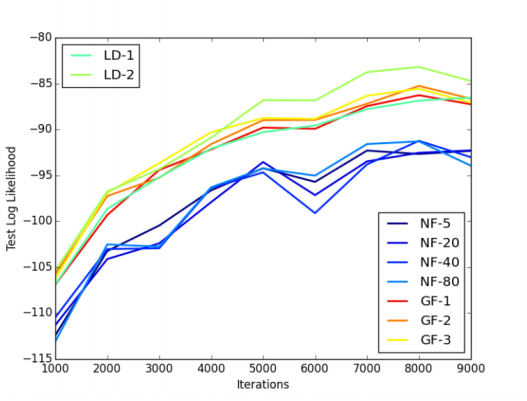
$$\sum_{\tau=1}^{\infty} \beta_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \beta_\tau^2 < \infty.$$

Утверждение [Welling, 2011]. Распределине $q^\tau(\boldsymbol{w})$ сходится к апостериорному распределению $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{f})$.
Изменение энтропии с учетом добавленного шума:

$$\hat{\mathrm{S}}\big(q^\tau(\boldsymbol{w})\big) \geq \frac{1}{2}|\boldsymbol{w}| \log\big(\exp(\frac{2\mathrm{S}(q^\tau(\boldsymbol{w}))}{|\boldsymbol{w}|}) + \exp(\frac{2\mathrm{S}(\epsilon)}{|\boldsymbol{w}|})\big).$$

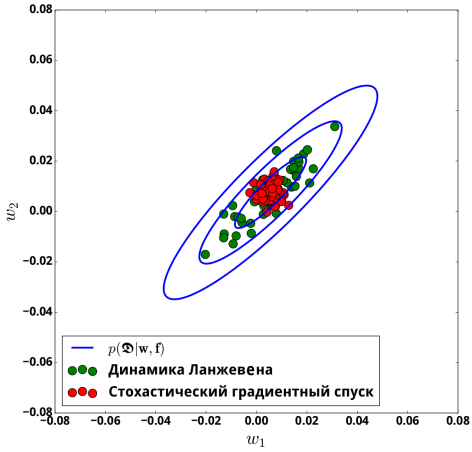
Стохастическая динамика Ланжевена в генеративных моделях

Altieri et al., 2015: будем сэмплировать скрытую переменную $\boldsymbol{z}$ и приближать его распределение к максимуму вариационной оценки с использованием динамики Ланжевена.



Стохастическая динамика Ланжевена

Распределение параметров после 2000 итераций:



Reparametrization trick: проблемы

Идея репараметризации:

$$\varepsilon = S_{\boldsymbol{\theta}}(\boldsymbol{w}), \quad \boldsymbol{w} = S_{\boldsymbol{\theta}}^{-1}(\varepsilon).$$

Тогда:

$$\nabla \boldsymbol{\theta} E_q f(\boldsymbol{w}) = \mathrm{E}_q \nabla_{\boldsymbol{\theta}} f(S_{\boldsymbol{\theta}}^{-1}(\varepsilon)) = \mathrm{E}_q \nabla_{\boldsymbol{w}} f(S_{\boldsymbol{\theta}}^{-1}(\varepsilon)) \nabla_{\boldsymbol{\theta}} S^{-1}(\varepsilon).$$

Пример:

$$w \sim \mathcal{N}(\mu, \sigma^2) \to S(w) = \frac{w - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Проблема: часто получение S^{-1} вычислительно дорого.

Implicit reparametrization trick

$$\nabla_{\boldsymbol{\theta}} E_q f(\boldsymbol{w}) = \mathrm{E}_q \nabla_{\boldsymbol{w}} f(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} \boldsymbol{w}.$$

Применим формулу полной производной к $\varepsilon = S_{\boldsymbol{\theta}}(\boldsymbol{w})$:

$$\nabla_{\boldsymbol{w}} S_{\boldsymbol{\theta}}(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} \boldsymbol{w} + \nabla_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}(\boldsymbol{w}) = 0 \rightarrow$$

$$\rightarrow \nabla_{\boldsymbol{\theta}} \boldsymbol{w} = -(\nabla_{\boldsymbol{w}} S_{\boldsymbol{\theta}}(\boldsymbol{w}))^{-1} \nabla_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}.$$

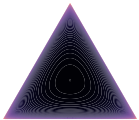
Получили формулу без обратной функции к S.
Для одномерных пространств в качестве S можно взять, например:

$$S(\boldsymbol{w}) = F(\boldsymbol{w}|\boldsymbol{\theta}) \sim \mathcal{U}(0, 1).$$

Table 4: Test negative log-likelihood (lower is better) for VAE on MNIST. Mean $\pm$ standard deviation over 5 runs. The von Mises-Fisher results are from [9].

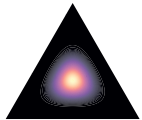
Prior	Variational posterior	$D = 2$	$D = 5$	$D = 10$	$D = 20$	$D = 40$
$\mathcal{N}(0, 1)$	$\mathcal{N}(\mu, \sigma^2)$	131.1 ± 0.6	107.9 ± 0.4	92.5 ± 0.2	88.1 ± 0.2	88.1 ± 0.0
$\text{Gamma}(0.3, 0.3)$	$\text{Gamma}(\alpha, \beta)$	132.4 ± 0.3	108.0 ± 0.3	94.0 ± 0.3	90.3 ± 0.2	90.6 ± 0.2
$\text{Gamma}(10, 10)$	$\text{Gamma}(\alpha, \beta)$	135.0 ± 0.2	107.0 ± 0.2	92.3 ± 0.2	88.3 ± 0.2	88.3 ± 0.1
$\text{Uniform}(0, 1)$	$\text{Beta}(\alpha, \beta)$	128.3 ± 0.2	107.4 ± 0.2	94.1 ± 0.1	88.9 ± 0.1	88.6 ± 0.1
$\text{Beta}(10, 10)$	$\text{Beta}(\alpha, \beta)$	131.1 ± 0.4	$\mathbf{106.7} \pm 0.1$	$\mathbf{92.1} \pm 0.2$	$\mathbf{87.8} \pm 0.1$	$\mathbf{87.7} \pm 0.1$
$\text{Uniform}(-\pi, \pi)$	$\text{vonMises}(\mu, \kappa)$	$\mathbf{127.6} \pm 0.4$	107.5 ± 0.4	94.4 ± 0.5	90.9 ± 0.1	91.5 ± 0.4
$\text{vonMises}(0, 10)$	$\text{vonMises}(\mu, \kappa)$	130.7 ± 0.8	107.5 ± 0.5	92.3 ± 0.2	$\mathbf{87.8} \pm 0.2$	87.9 ± 0.3
$\text{Uniform}(S^D)$	$\text{vonMisesFisher}(\boldsymbol{\mu}, \kappa)$	132.5 ± 0.7	108.4 ± 0.1	93.2 ± 0.1	89.0 ± 0.3	90.9 ± 0.3

Дискретные распределения: релаксация

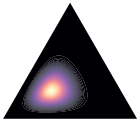


$\bar{\alpha} = [1, 1, 1], t = 0.9$

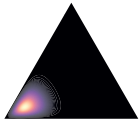
$t = 1.0$



$t = 10.0$



$\bar{\alpha} = [0.5, 0.25, 0.25], t = 30.0$



$[0.75, 0.125, 0.125]$

$[0.9, 0.05, 0.05]$

Дискретные распределения в вариационном выводе

Релаксация:

- Распределение Дирихле (если использовать Implicit reparametrization trick)
- Gumbel-softmax:

$$p(\boldsymbol{w}) = \Gamma(k)\tau^{k-1}\left(\sum_{i=1}^{k} \alpha_i / w_i\right)^{-k} \prod_{i=1}^{k} (\alpha_i / w_i \tau + 1)$$

  - Возможна репараметризация
  - нет аналитической формы для KL
- Invertible Gaussian reparametrization:

$$p(\boldsymbol{w}) = \bar{\text{softmax}}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \sim \mathcal{N},$$

(в знаменатель softmax добавляется константа для гарантии обратимости функции)

  - Возможна репараметризация
  - $KL(\boldsymbol{w}_1|\boldsymbol{w}_2) = KL(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2)$
  - Интерпретация параметров сильно сложнее, чем у GS или Дирихле

Локальная репараметризация

Пусть $y = \text{ReLU}(\boldsymbol{X}\boldsymbol{W})$ и матрица параметров $\boldsymbol{W}$ распределена нормально: $w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$.
Тогда результатом линейной операции $\boldsymbol{X}\boldsymbol{W}$ будет гауссовая матрица:

$$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{W}, \quad G_{i,j} \sim \mathcal{N}(\sum_k x_{i,k}\mu_{k,j}, \sum_k x^2_{i,k}\sigma^2_{k,j}).$$

Вместо сэмлирования полноценного вектора параметров для каждого элемента батча на каждом шаге оптимизации, сэмплируем элементы из $\boldsymbol{G}$ (то, что идет перед ReLU).

Если $\boldsymbol{w}$ принимает значения из дискретного набора значений, то пользуясь ЦПТ (в формулировке Ляпунова):

$$\sum_k x_{i,k} w_{k,j} \sim \mathcal{N}(\cdot, \cdot).$$

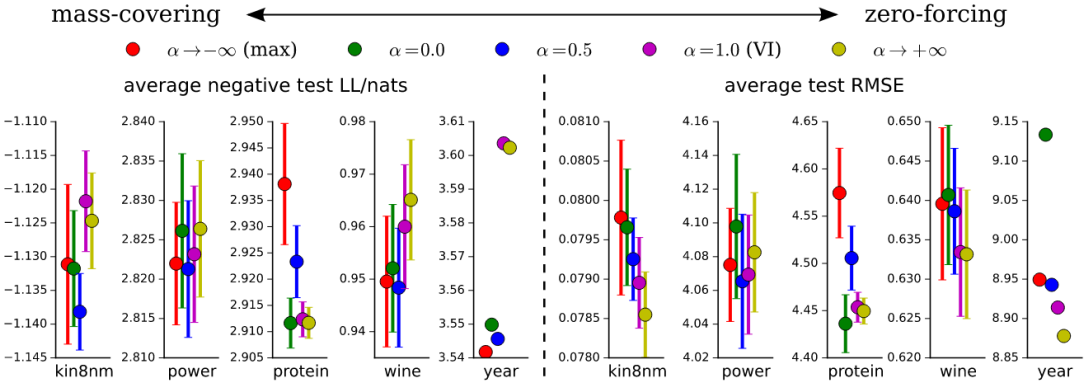
Значит, к дискретным значения также можно применить локальную репараметризацию.

Дивергенция Реньи

$$D_\alpha(p(\boldsymbol{w})|q(\boldsymbol{w})) = \frac{1}{\alpha - 1} \log \int p(\boldsymbol{w})^\alpha q(\boldsymbol{w})^{1-\alpha} d\boldsymbol{w}.$$

Table 1: Special cases in the Rényi divergence family.

α	Definition	Notes
$\alpha \to 1$	$\int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$	*Kullback-Leibler (KL) divergence*, used in VI ($\mathrm{KL}[q\|\|p]$) and EP ($\mathrm{KL}[p\|\|q]$)
$\alpha = 0.5$	$-2\log(1 - \mathrm{Hel}^2[p\|\|q])$	function of the square *Hellinger distance*
$\alpha \to 0$	$-\log \int_{p(\boldsymbol{\theta})>0} q(\boldsymbol{\theta}) d\boldsymbol{\theta}$	zero when $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$ (not a divergence)
$\alpha = 2$	$-\log(1 - \chi^2[p\|\|q])$	proportional to the χ^2-divergence
$\alpha \to +\infty$	$\log \max_{\boldsymbol{\theta} \in \Theta} \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$	*worst-case regret* in *minimum description length principle*


mass-covering
zero-forcing
$\alpha \to -\infty$ (max)
$\alpha = 0.0$
$\alpha = 0.5$
$\alpha = 1.0$ (VI)
$\alpha \to +\infty$
average negative test LL/nats
average test RMSE
kin8nm
power
protein
wine
year

Литература и прочие ресурсы

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Salimans, Tim, Diederik Kingma, and Max Welling, 2015. Markov chain monte carlo and variational inference: Bridging the gap
- Altieri: http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf
- Stephan Mandt, Matthew D. Hoffman, David M. Blei, 2017. Stochastic Gradient Descent as Approximate Bayesian Inference
- Бахтеев О. Ю., Стрижов В. В. Выбор моделей глубокого обучения субоптимальной сложности //Автоматика и телемеханика. – 2018. – №. 8. – С. 129-147.
- Figurnov M., Mohamed S., Mnih A. Implicit reparameterization gradients //arXiv preprint arXiv:1805.08498. – 2018.
- Jang E., Gu S., Poole B. Categorical reparameterization with gumbel-softmax //arXiv preprint arXiv:1611.01144. – 2016.
- Potapczynski A., Loaiza-Ganem G., Cunningham J. P. Invertible gaussian reparameterization: Revisiting the gumbel-softmax //arXiv preprint arXiv:1912.09588. – 2019.
- Maddison C. J., Mnih A., Teh Y. W. The concrete distribution: A continuous relaxation of discrete random variables //arXiv preprint arXiv:1611.00712. – 2016.
- Shayer O., Levi D., Fetaya E. Learning discrete weights using the local reparameterization trick //arXiv preprint arXiv:1710.07739. – 2017.
- Li Y., Turner R. E. Rényi Divergence Variational Inference //arXiv preprint arXiv:1602.02311. – 2016.