Bayesian dark knowledge[1]

Андрей Атанов[1]

[1]Высшая Школа Экономики

Москва, 2017

[1]Anoop Korattikara Balan и др. (2015). “Bayesian Dark Knowledge”. В: *CoRR* abs/1506.04416. URL: http://arxiv.org/abs/1506.04416.

Байесовский вывод

- Задание параметрической модели $p(y|x,\theta)$ и априорного распределения на параметры $p(\theta)$
- Наблюдение данных $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, вывод апостериорного распределения:

$$p(\theta|\mathcal{D}_n) = \frac{p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)}{p(\mathcal{D}_n)}$$

- Для нового объекта x^* вывод posterior predictive distribution:

$$p(y^*|x^*,\mathcal{D}_n) = \int p(y^*|x^*,\theta)p(\theta|\mathcal{D}_n)d\theta$$

Методы аппроксимации:

- MAP оценка
- Приближение апостериорного распределения полностью факторизованнным распределением
- ELBO + reparametrization trick

Недостатки существующих методов

- Плохая аппроксимация в случае многомодальных апостериорных распределений
- В тестовой фазе необходимо сэмплировать большое количество моделей
- Требуют большее количество памяти чем методы использующие точечные оценки
- Могут быть трудны в реализации и переносе на более сложные архитектуры

MCMC

$$p(\theta|\mathcal{D}_N) \approx \frac{1}{S}\sum_{s=1}^{S}\delta(\theta - \theta^s)$$

$$p(y^*|x^*, \mathcal{D}_n) = \mathbb{E}_{\theta|\mathcal{D}_n} p(y^*|x^*, \theta) \approx \frac{1}{S}\sum_{s=1}^{S} p(y^*|x^*, \theta^s)$$

+ Более точная аппроксимация, чем точечные оценки
+ Простая реализация даже для сложных моделей
− $S \gg 1$
− В S раз (!) медленнее и больше требуемый объем памяти

Аппроксимация байесовского ансамблирования[2]

- TNN — ансамбль $T(y|x,\theta^s)$ из апостериорного распределения $p(\theta|\mathcal{D}_N)$ – Teacher NN.
- SNN — Student NN $S(y|x,w)$, аппроксимирующая ансамбль

$$KL(p(y|x,\mathcal{D}_n)||S(y|x,w)) \propto -\int p(y|x,\mathcal{D}_n)\log S(y|x,w)dy$$

$$= -\int\left[\int p(y|x,\theta)p(\theta|\mathcal{D}_n)d\theta\right]\log S(y|x,w)dy$$

$$= -\int p(\theta|\mathcal{D}_n)\left[\int p(y|x,\theta)\log S(y|x,w)dy\right]d\theta$$

$$\approx -\frac{1}{|\Theta|}\sum_{\theta^s\in\Theta}\mathbb{E}_{p(y|x,\theta^s)}\log S(y|x,w),\ \theta^s\sim p(\theta|\mathcal{D}_N)$$

[2] Anoop Korattikara Balan и др. (2015). "Bayesian Dark Knowledge". В: *CoRR* abs/1506.04416. URL: http://arxiv.org/abs/1506.04416.

Аппроксимация байесовского ансамблирования

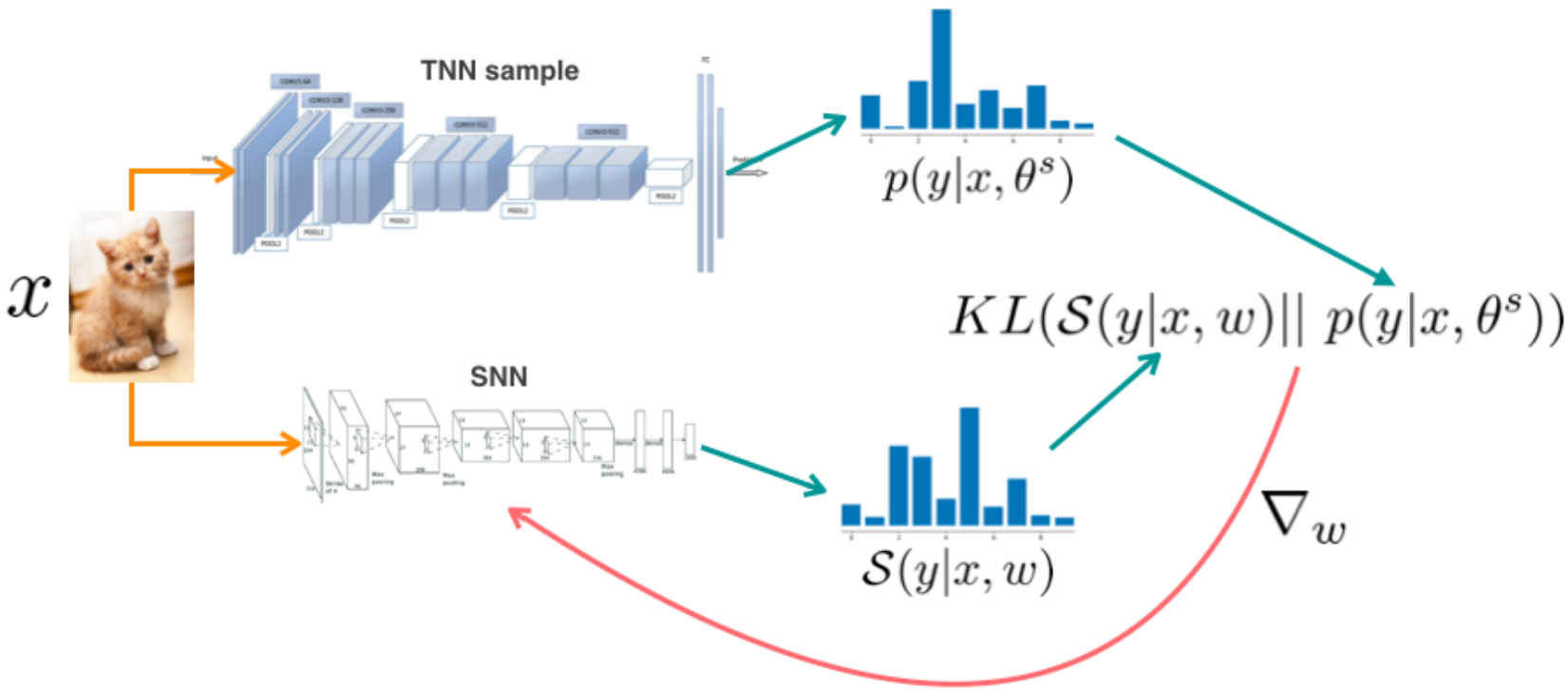
Функция потерь

$$L(w|x) = \frac{1}{|\Theta|} \sum_{\theta^s \in \Theta} \mathbb{E}_{p(y|x,\theta^s)} \log S(y|x,w)$$

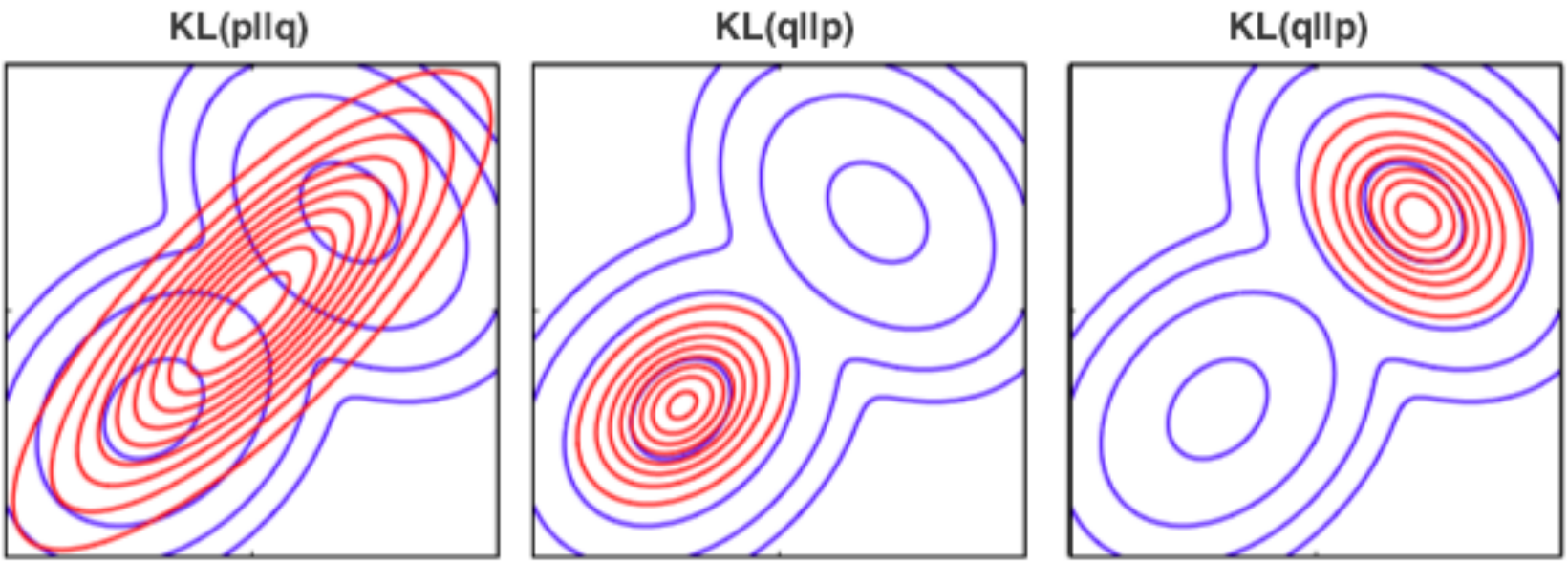
$$L(w) = \int L(w|x)p(x)dx \approx -\frac{1}{|\Theta|}\frac{1}{|\mathcal{D}'|} \sum_{\theta^s \in \Theta} \sum_{x \in \mathcal{D}'} \mathbb{E}_{p(y|x',\theta^s)} \log S(y|x',w)$$

1. $L(w)$ — MC оценка $\int p(x)KL(p||S)dx$ – достаточно точная при $|\Theta| \gg 1, |\mathcal{D}'| \gg 1$
2. Имеет форму $\sum\sum L(w|\theta^s, x')$. Можно воспользоваться SGD!
3. Можно семплировать θ^s с помощью SGLD или Dropout
4. примеры $\mathcal{D}'$ могут быть неразмеченными

Distillation[3]



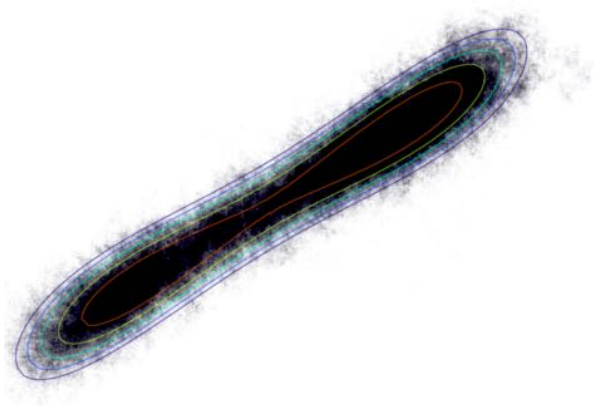
[3] Jeff Dean Geoffrey Hinton Oriol Vinyals (2014). "Distilling the Knowledge in a Neural Network". B:

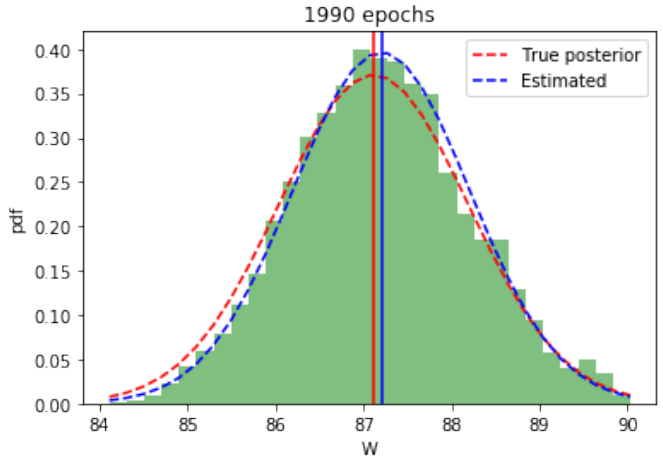

KL(p||q)
KL(q||p)
KL(q||p)

Stochastic gradient Langevin dynamics[4]

$$\theta_{t+1} = \theta_t + \frac{\eta_t}{2}\left(\nabla_\theta \log p(\theta) + \frac{N}{M}\sum_{(x,y)\in\mathcal{D}_M} \log p(y|x,\theta)\right) + \mathcal{N}(0,\eta_t)$$

- использует градиент апостериорного распределения для определения направления
- добавляет нормальный шум на каждом шаге, чтобы не сойтись к MAP оценке





[4] Max Welling и Yee W Teh (2011). "Bayesian learning via stochastic gradient Langevin dynamics". В: с. 681—688.

Алгоритм

Algorithm 1 Distil Ensemble via SGLD

for t = 1 **to** T **do**

Sample minibatch $\mathcal{D}_M$ from train data set.

Sample $z_t \sim \mathcal{N}(0, \eta_t I)$

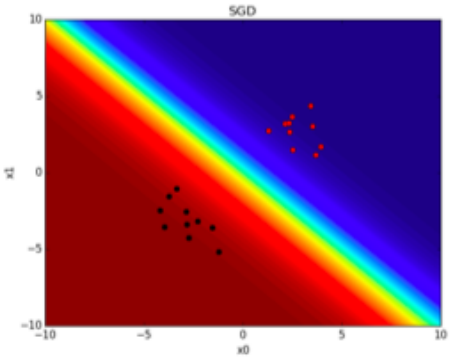
$$\theta_{t+1} = \theta_t + \frac{\eta_t}{2} \nabla_\theta \log(\theta|\mathcal{D}_M) + z_t$$

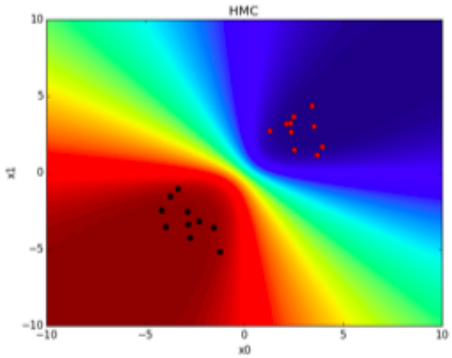
Sample $\mathcal{D}'$ by student data generator

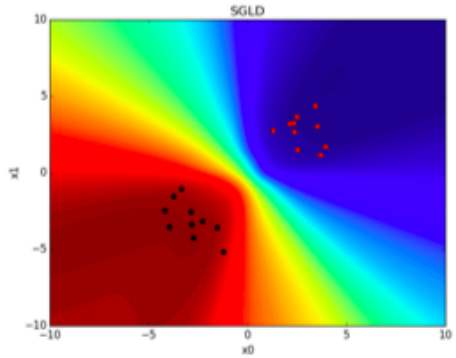
$$w_{t+1} = w_t - \xi_t \left(\frac{1}{|\mathcal{D}'|} \sum_{x' \in \mathcal{D}'} \nabla_w L(w|\theta_{t+1}, x') \right)$$

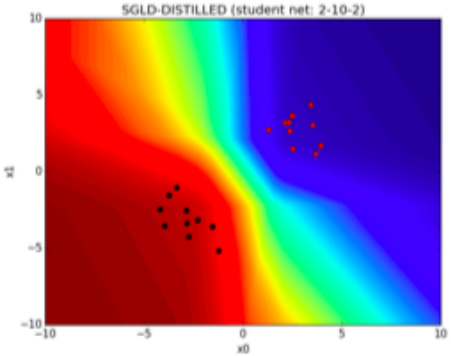
end for

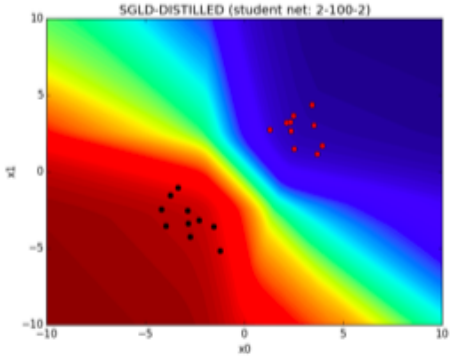
Toy 2-d problem

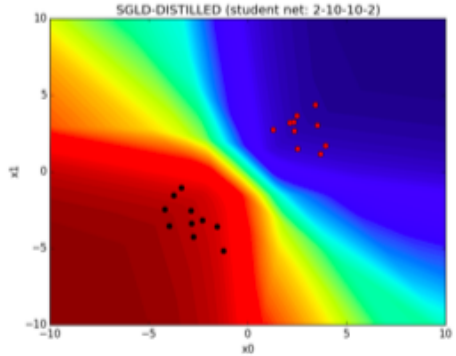








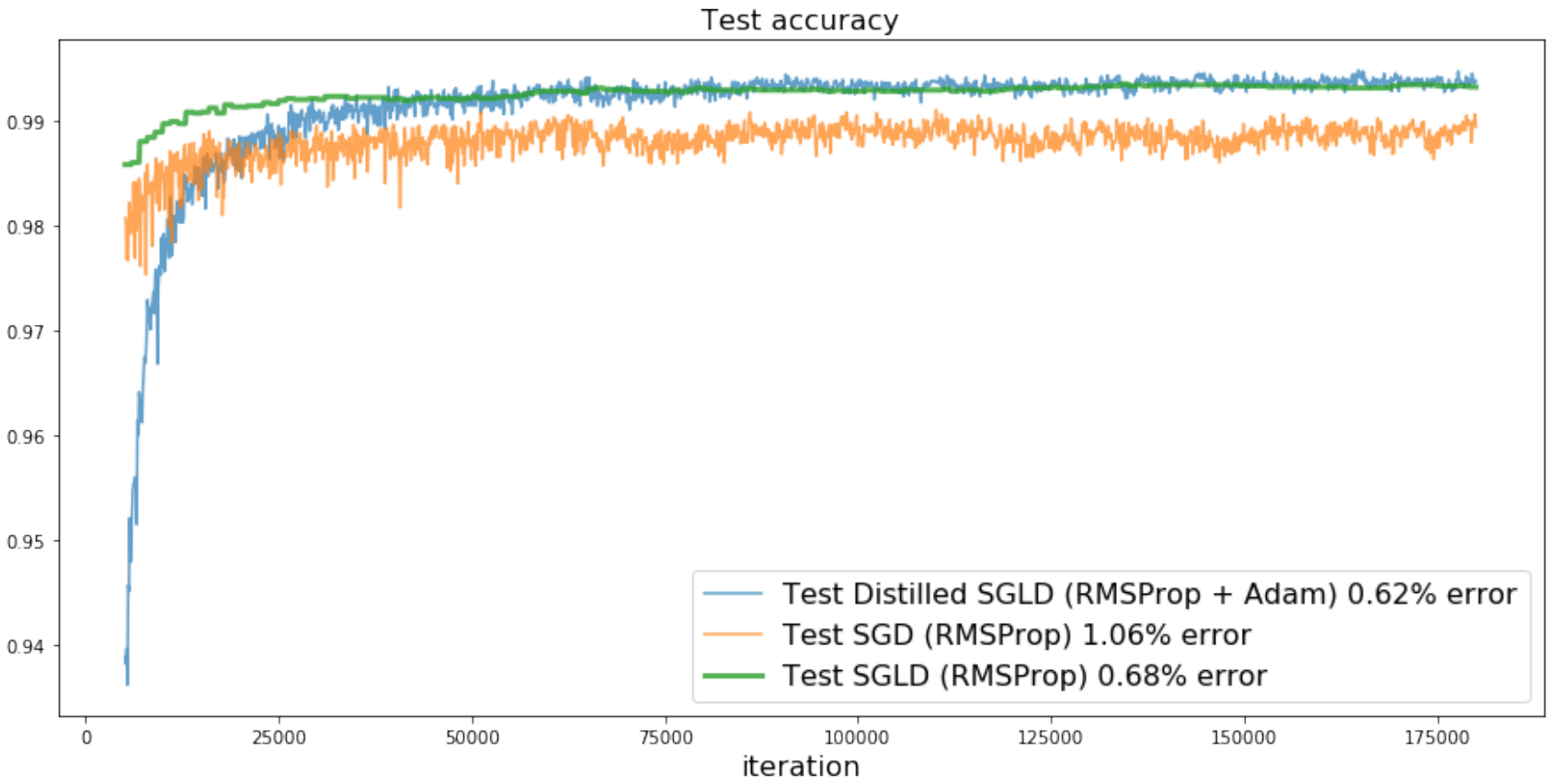




Results on SGLD

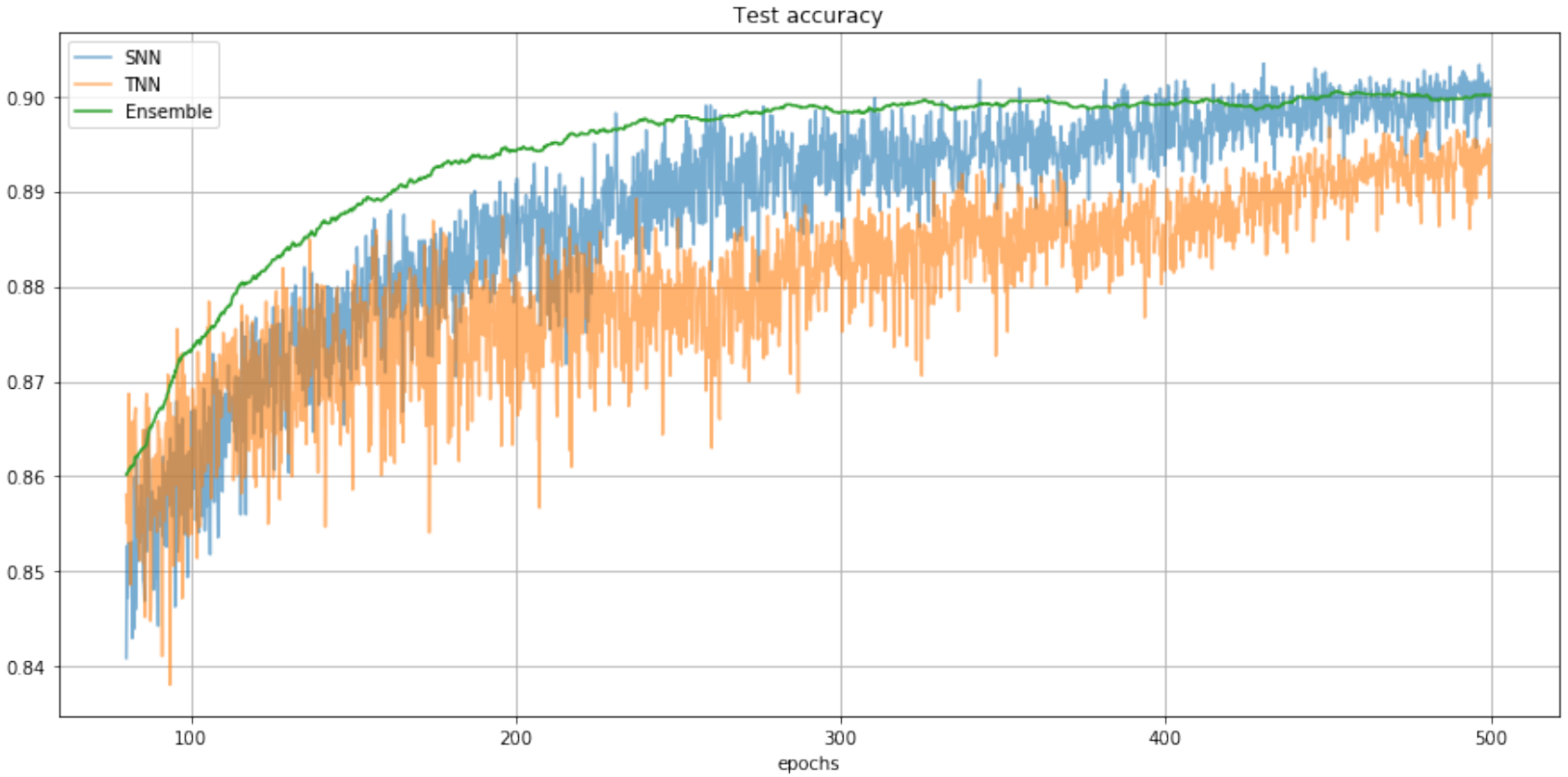
MNIST LeNet

SNN приближает ансамбль, полученный с помощью SGLD



Results on SGLD

CIFAR-10 VGG19



Results on Dropout

CIFAR-10 VGG19