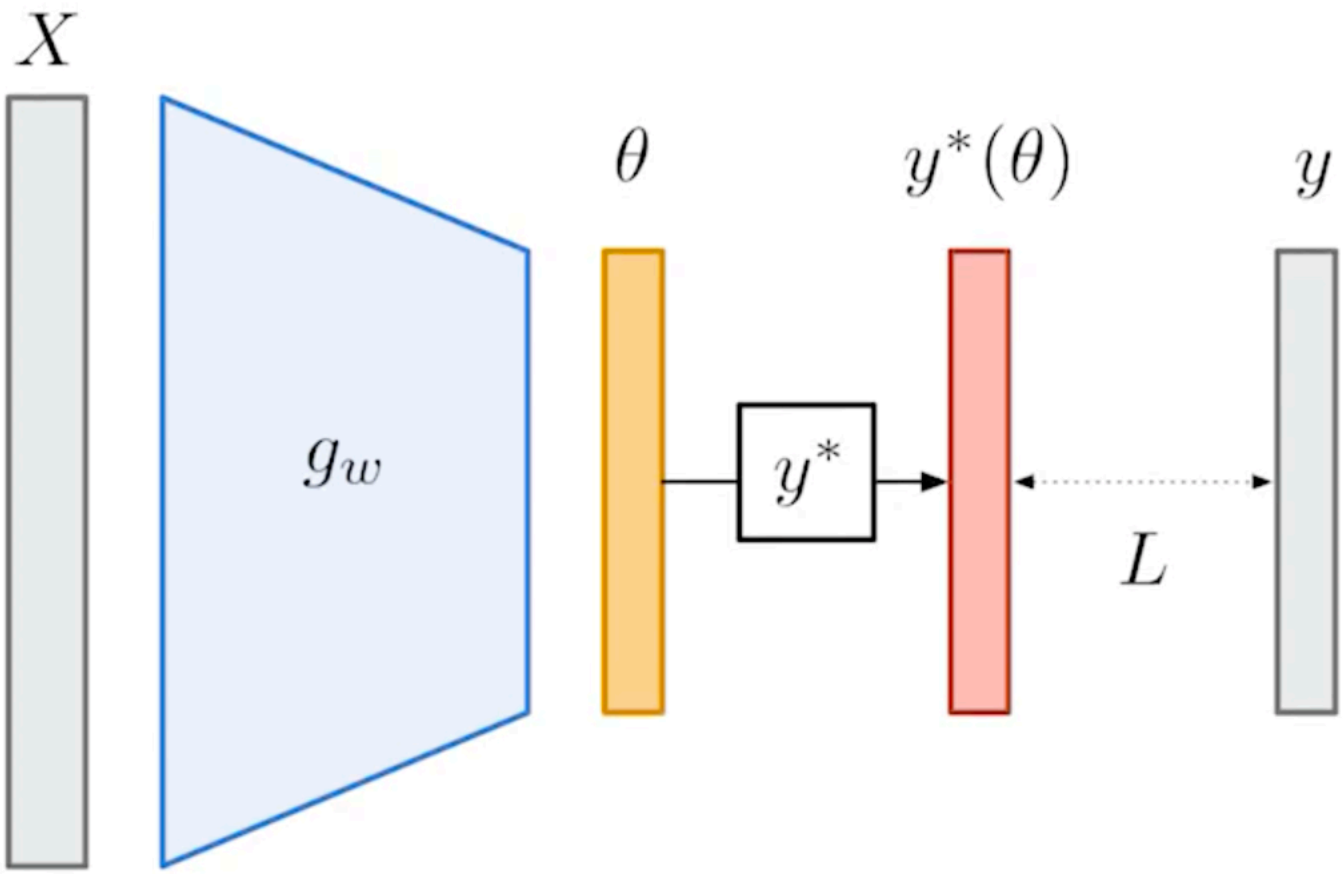
Левина Александра БПМИ171

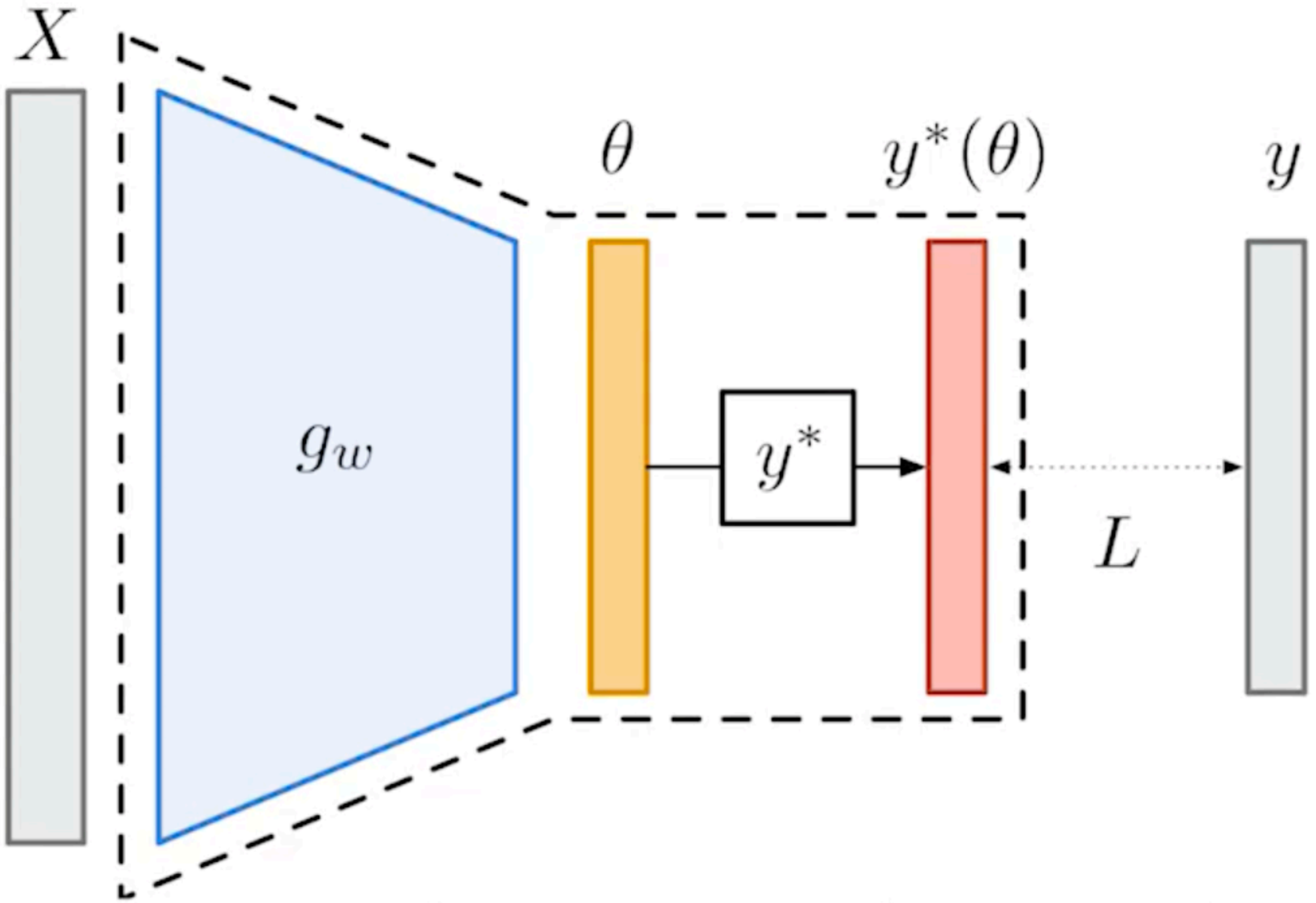
Learning with Differentiable Perturbed Optimizers

Обучение с дифференцируемыми возмущенными оптимизаторами

Дискретные задачи



Дискретные задачи



Возмущенный максимизатор

Задача дискретной оптимизации

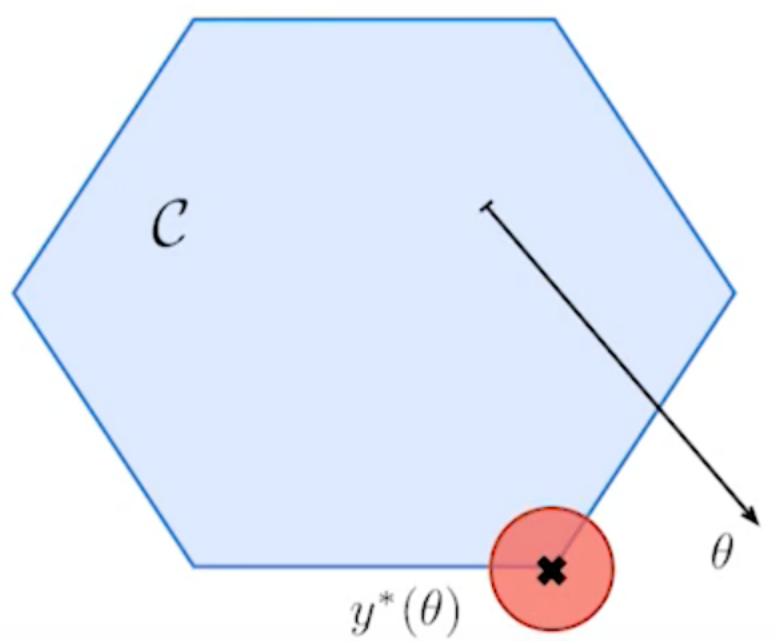
$Y \in \mathbb{R}^d$

$\mathcal{C}$ – выпуклая оболочка Y

$\theta \in \mathbb{R}^d$

$$F(\theta) = \max_{y \in \mathcal{C}} \langle y, \theta \rangle$$

$$y^*(\theta) = \arg\max_{y \in \mathcal{C}} \langle y, \theta \rangle$$

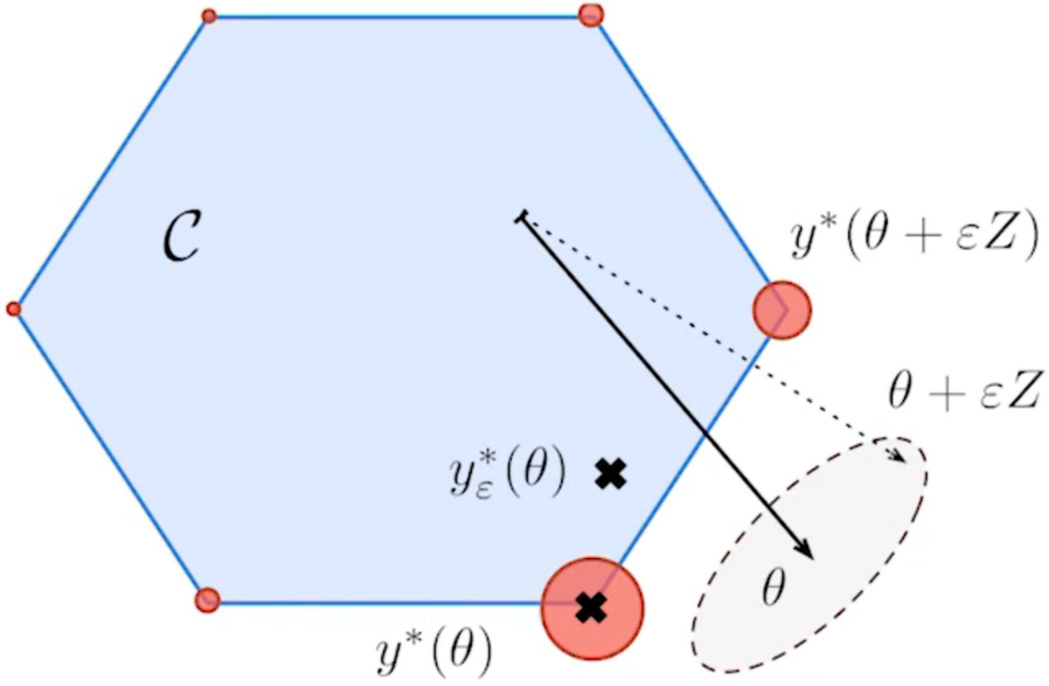


Возмущенный максимизатор

$\epsilon > 0$

Z с положительной и дифференцируемой плотностью

p_θ для $y \in Y : p_\theta(y) = P(y^*(\theta + \epsilon Z) = y)$



$$F_\varepsilon(\theta) = \mathbf{E}[F(\theta + \varepsilon Z)] = \mathbf{E}[\max_{y \in \mathcal{C}} \langle y, \theta + \varepsilon Z \rangle]$$

$$y^*_\varepsilon(\theta) = \mathbf{E}_{p_\theta(y)}[Y] = \mathbf{E}[\arg\max_{y \in \mathcal{C}} \langle y, \theta + \varepsilon Z \rangle] = \mathbf{E}[\nabla_\theta \max_{y \in \mathcal{C}} \langle y, \theta + \varepsilon Z \rangle] = \nabla_\theta F_\varepsilon(\theta)$$

Свойства возмущенного максимизатора

$\epsilon\Omega$ – двойственная функции F_ϵ

$$y^*_\varepsilon(\theta) = \arg\max_{y \in \mathcal{C}} \left\{ \langle y, \theta \rangle - \varepsilon\,\Omega(y) \right\}$$

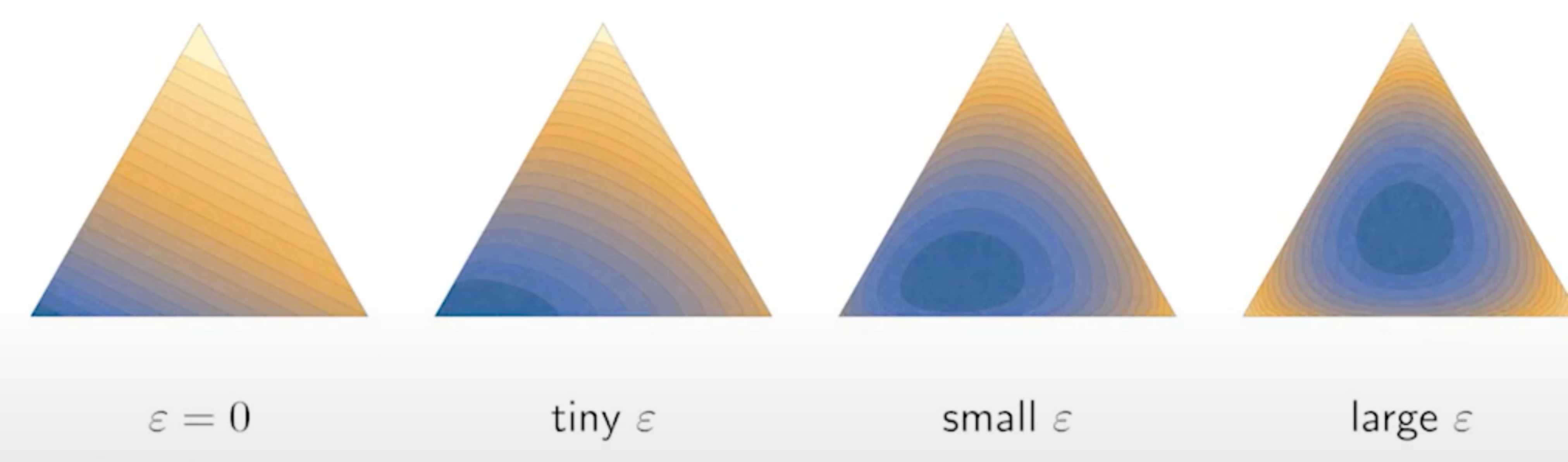
Свойства возмущенного максимизатора

Регуляризация

$$y^*_\varepsilon(\theta) = \underset{y \in \mathcal{C}}{\arg\max} \left\{ \langle y, \theta \rangle - \varepsilon\, \Omega(y) \right\}$$

$$\epsilon \to 0 \Rightarrow y^*_\epsilon(\theta) \to y^*(\theta)$$

$$\epsilon \to \infty \Rightarrow y^*_\epsilon(\theta) \to \mathrm{argmin}_y \Omega(y)$$



Свойства возмущенного максимизатора

Строгая выпуклость

F_ϵ – строго выпуклая

$y^*_\epsilon(\theta)$ – везде локально неконстантный

Свойства возмущенного максимизатора

Z с положительной и дифференцируемой плотностью $d\mu(z) \propto \exp(-\nu(z))dz$

Простые мат ожидания

$$F_\varepsilon(\theta) = \mathbf{E}[F(\theta + \varepsilon Z)]$$

$$y^*_\varepsilon(\theta) = \nabla_\theta F_\varepsilon(\theta) = \mathbf{E}[y^*(\theta + \varepsilon Z)] = \mathbf{E}[F(\theta + \varepsilon Z)\nabla_z \nu(Z)/\varepsilon]$$

$$J_\theta\, y^*_\varepsilon(\theta) = \mathbf{E}[y^*(\theta + \varepsilon Z)\nabla_z \nu(Z)^\top/\varepsilon] = \mathbf{E}[F(\theta + \varepsilon Z)(\nabla_z \nu(Z)\nabla_z \nu(Z)^\top - \nabla_z^2 \nu(Z))/\varepsilon^2]$$

Имплементация модели

Монте-Карло

Выпуклая задача оптимизации $$y^*_\varepsilon(\theta) = \operatorname*{arg\,max}_{y \in \mathcal{C}} \{\langle y, \theta \rangle - \varepsilon\, \Omega(y)\}$$

$(Z^{(1)}, \cdots, Z^{(m)})$ – независимые одинаково распределённые случайные величины из $\mu(z)$

$$y^{(m)} = y^*(\theta + \varepsilon Z^{(m)}) = \operatorname{arg\,max}_{y \in \mathcal{C}} \langle y, \theta + \varepsilon Z^{(m)} \rangle$$

Оценка Монте-Карло для $y^*_\varepsilon(\theta)$ $$\bar{y}_{\varepsilon, M}(\theta) = \frac{1}{M} \sum_{m=1}^{M} y^{(m)}$$

$$\mathbf{E}[y^{(m)}] = y^*_\epsilon(\theta)\ \forall\ m$$

$$p_\theta(y) = P(y^*(\theta + \epsilon Z) = y)$$

Функция потерь Фенхеля-Янга

Формула

$$L_\varepsilon(\theta\,; y) = F_\varepsilon(\theta) + \varepsilon\,\Omega(y) - \langle \theta, y \rangle$$

Формула градиента

$$\nabla_\theta L_\varepsilon(\theta\,; y) = \nabla_\theta F_\varepsilon(\theta) - y = y^*_\varepsilon(\theta) - y$$

минимум в $y^*_\varepsilon(\theta) = y$

для любого Y $\mathbf{E}[L_\varepsilon(\theta; Y)] = L_\varepsilon(\theta; \mathbf{E}[Y]) + C$

Функция потерь Фенхеля-Янга

$$L_\varepsilon(\theta\,;y) = F_\varepsilon(\theta) + \varepsilon\,\Omega(y) - \langle\theta, y\rangle \qquad \nabla_\theta L_\varepsilon(\theta\,;y) = \nabla_\theta F_\varepsilon(\theta) - y = y^*_\varepsilon(\theta) - y$$

Обучение с учителем

$$L_{\varepsilon,\text{emp}}(w) = \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(g_w(x_i)\,;y_i)$$

$$\nabla_w L_{\varepsilon,\text{emp}}(w) = \frac{1}{n}\sum_{i=1}^{n} J_w\, g_w(x_i)\cdot(y^*_\varepsilon(g_w(x_i)) - y_i)$$

$$\bar{\gamma}_{i,M}(w) = J_w\, g_w(x_i)\Big(\frac{1}{M}\sum_{m=1}^{M} y^*\big(g_w(x_i) + \varepsilon Z^{(m)}\big) - y_i\Big)$$

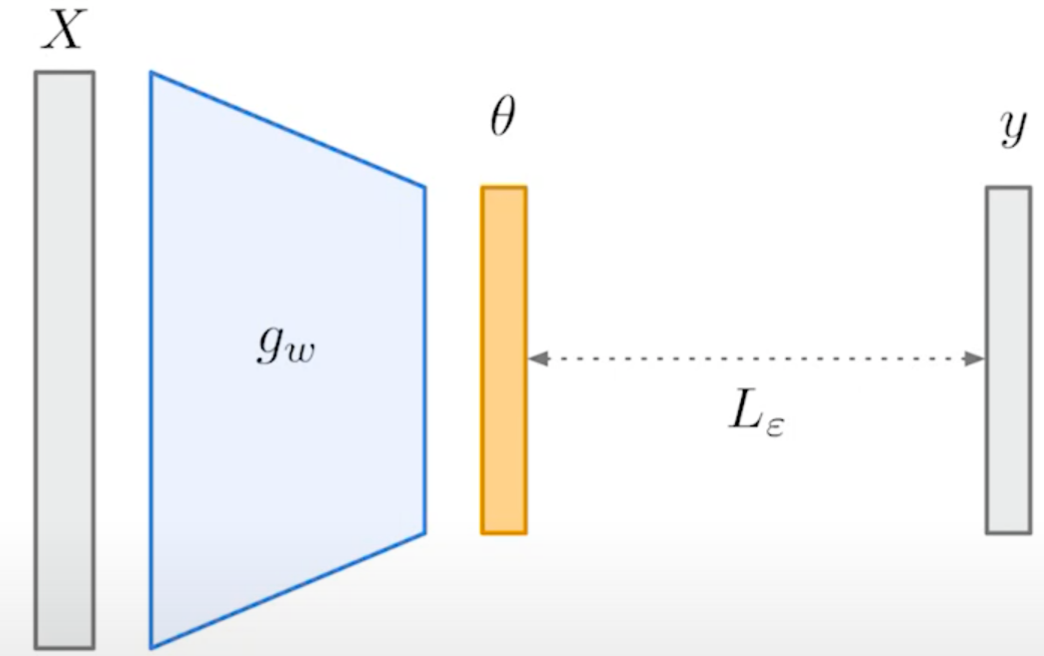
Итоги

- Предложен общий метод трансформации дискретных оптимизаторов, который применим к любой black-box модели

- Метод предлагает способ дифференцирования аргмакса с хорошо определенным Якобианом

- Все производные легко приблизить методом Монте-Карло, что дает вычислительную эффективность

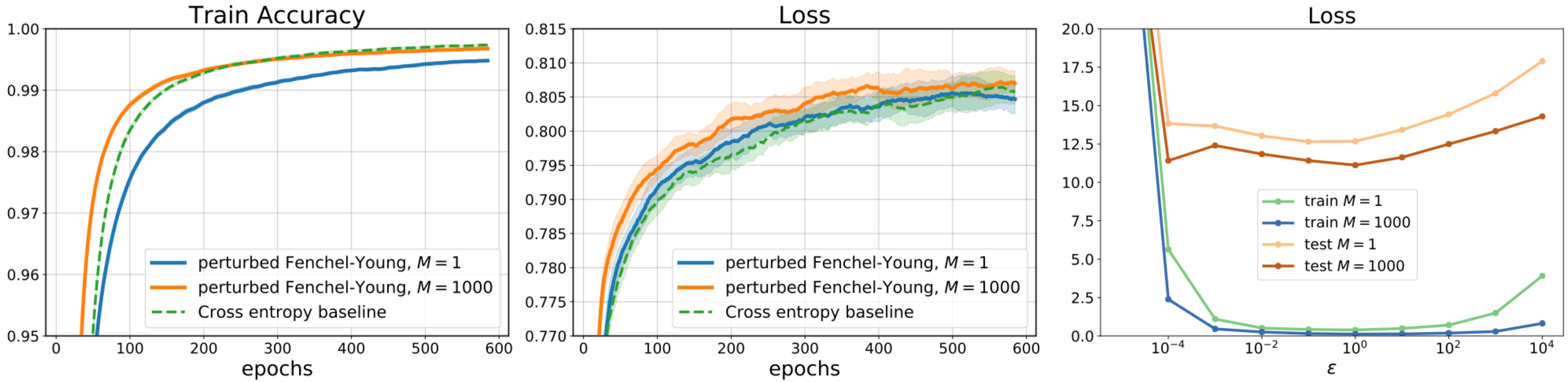
- Предложена удобная функция потерь



Эксперименты

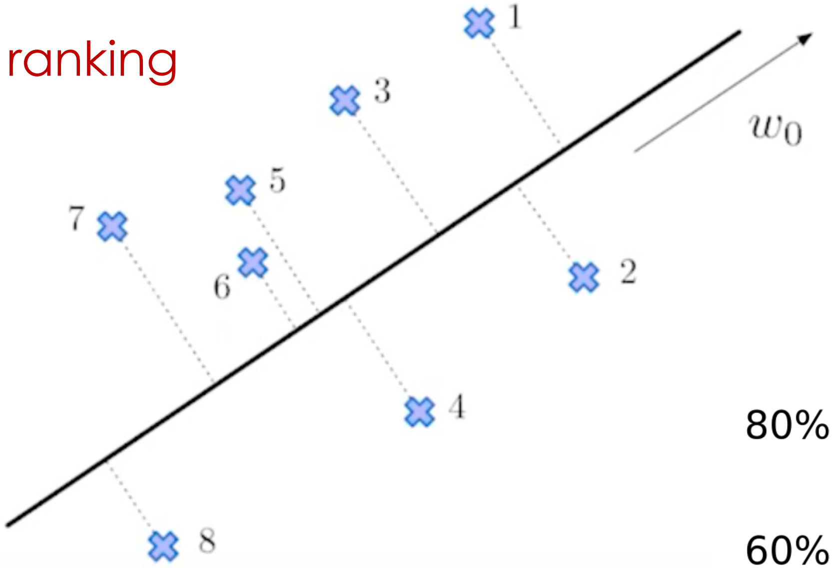
Классификация на 10 классов

- Возмущенный максимизатор с нормальным шумом
- Vanilla CNN (4 свертки + 2 полносвязных слоя)

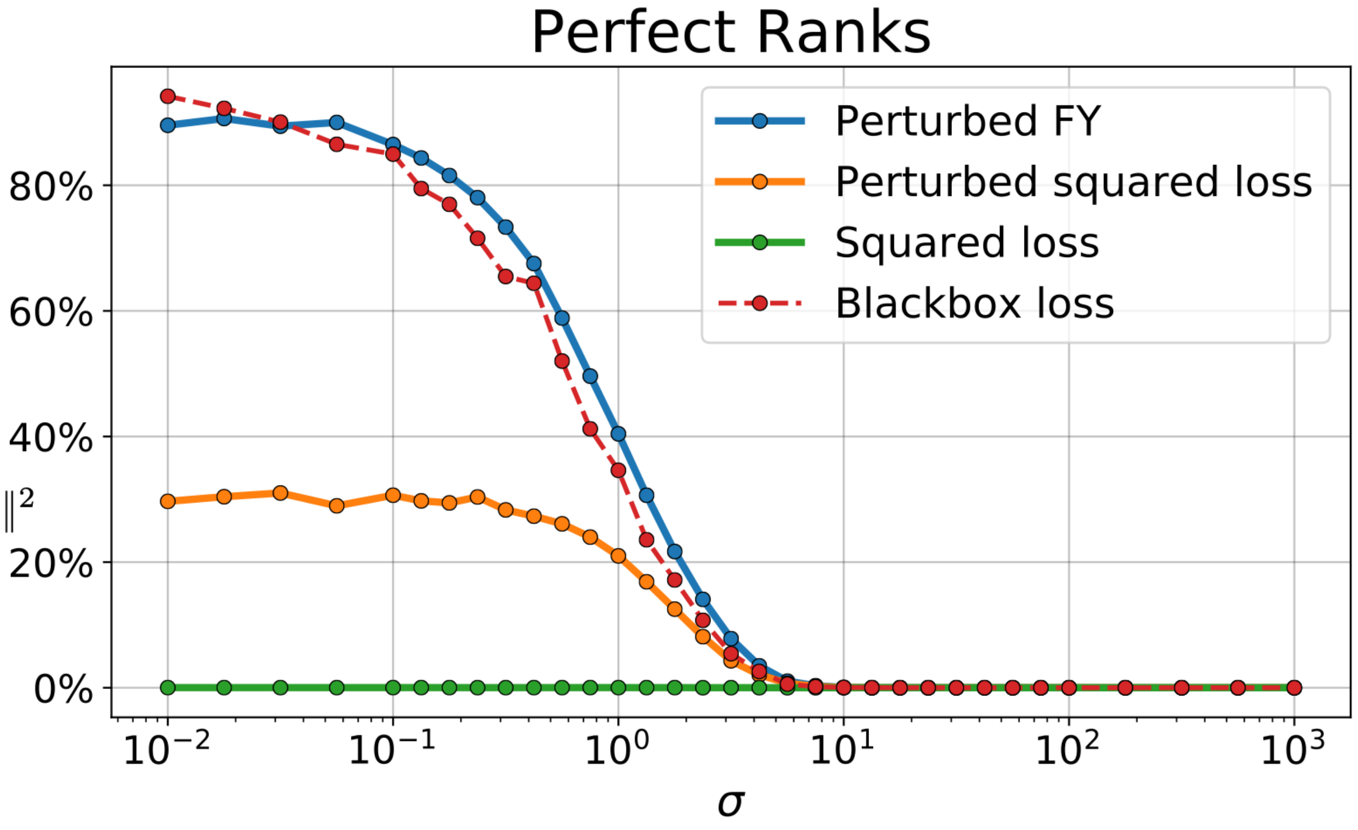


Эксперименты

Label ranking



$$y_i = \arg\max_y \langle x_i^\top w_0 + \sigma Z_i, y\rangle$$



Perturbed + Squared loss (proposed): $\frac{1}{2}\|y_i - y^*_\varepsilon(g_w(x_i))\|^2$

Squared loss: $\frac{1}{2}\|y_i - g_w(x_i)\|^2$

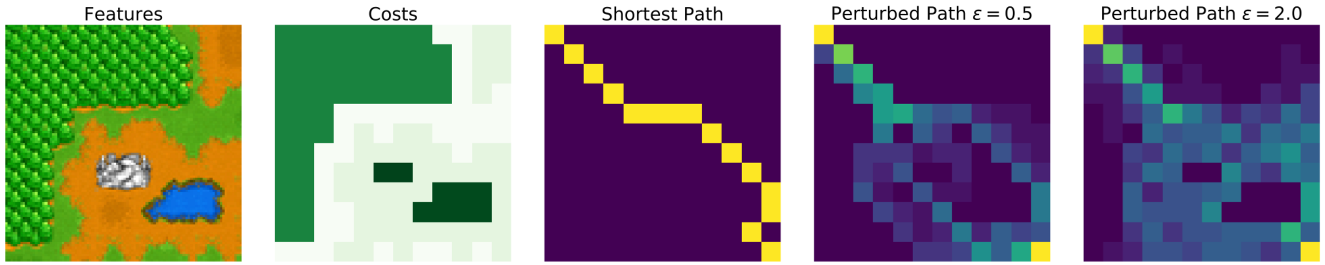
Blackbox loss: $\frac{1}{2}\|y_i - y^*(g_w(x_i))\|^2$

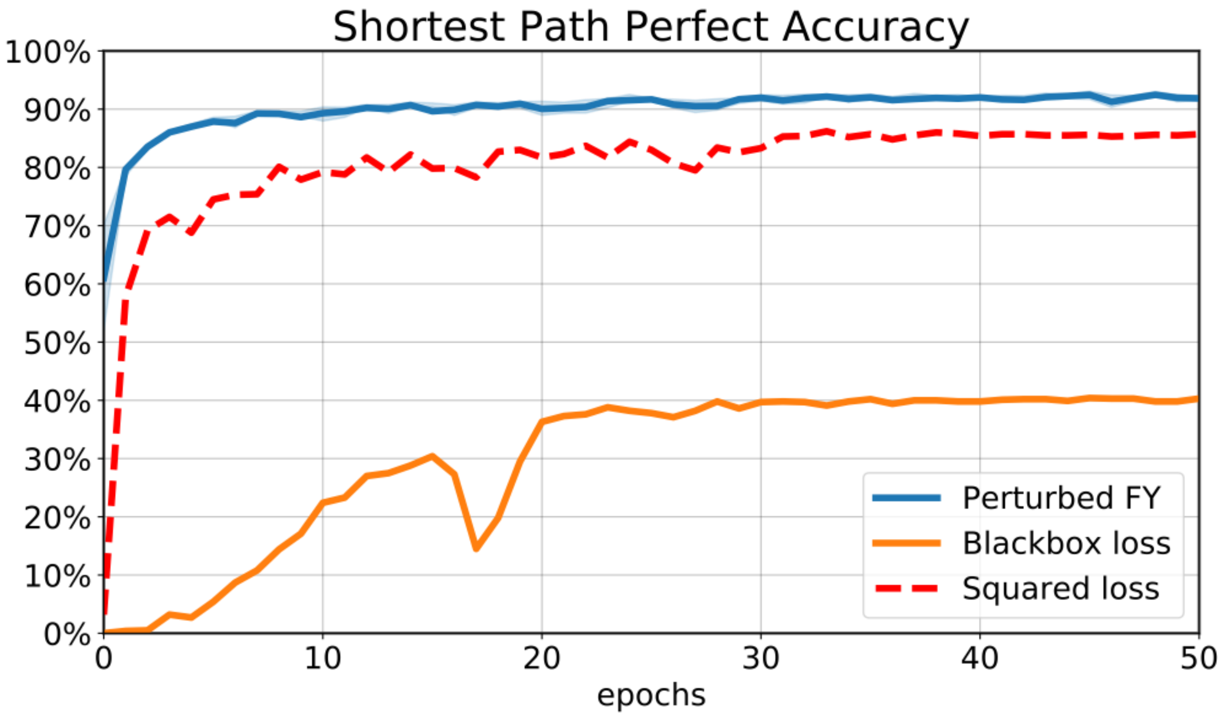
+ приближение градиента из статьи «Differentiation of blackbox combinatorial solvers» (2019)

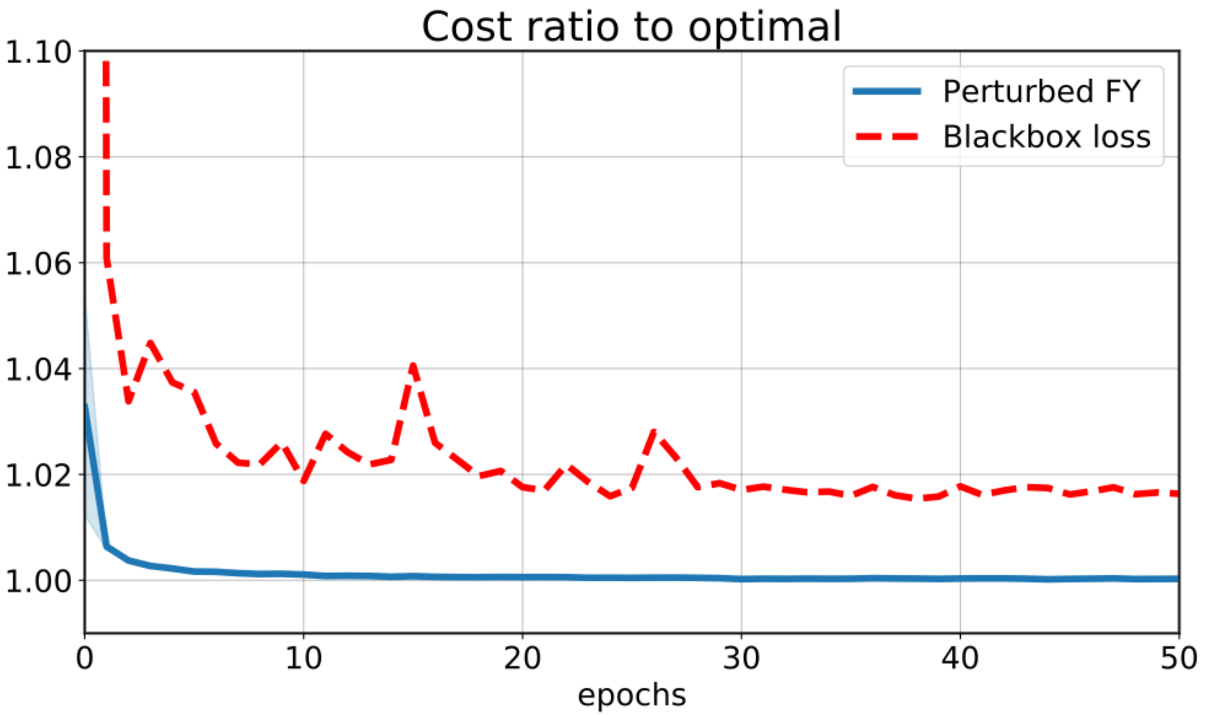
Эксперименты

Поиск кратчайшего пути

- Первые пять слоев ResNet18
- ε=1, M=1







Вопросы

1. Какая мотивация у использования возмущенных максимизаторов, какие проблемы они решают?

2. Напишите формулу функции потерь Фенхеля-Янга для возмущенного максимизатора и объясните все обозначения.

3. Какие свойства построенной модели описывают авторы в статье?