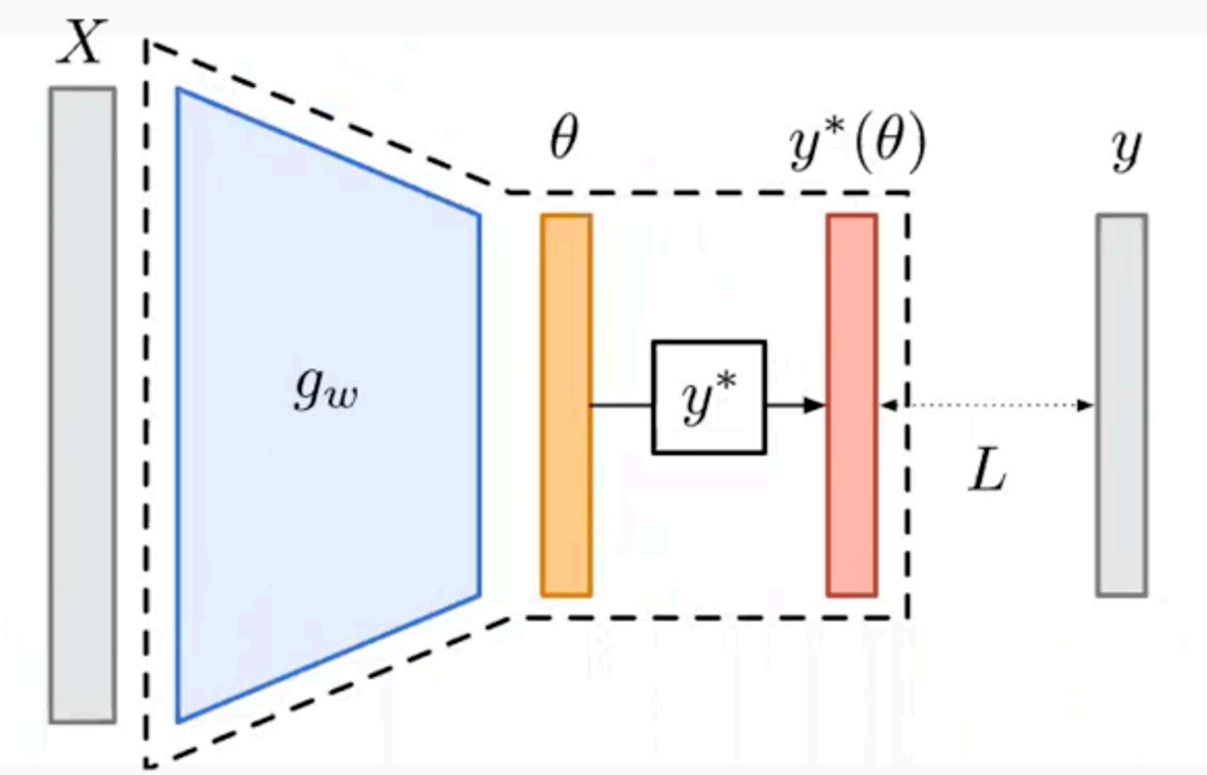
Learning with Differentiable Perturbed Optimizers

Обучение с дифференцируемыми возмущенными оптимизаторами

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, Francis Bach

Zakieva Azaliya, AMI172
HSE Research Seminar
December 2020

Проблема: дискретность



Примеры:

- Задача ранжирования (предсказания перестановок)
- Выбор ближайших соседей
- Поиск кратчайшего пути

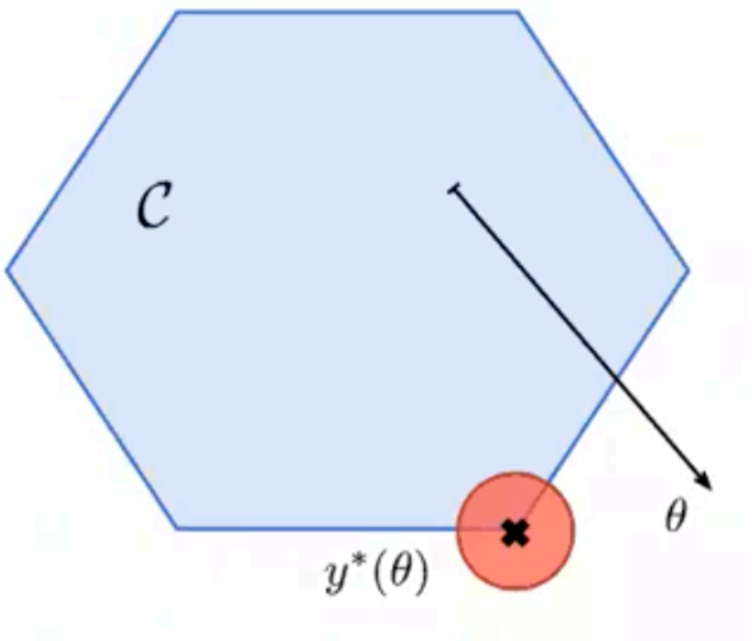
Возмущенный максимизатор (Perturbed maximizer)

Общая задача дискретной оптимизации:

$\mathcal{Y} \subset \mathbb{R}^d$ - конечное мн-во точек

$\mathcal{C}$ - выпуклая оболочка $\mathcal{Y}$

$\theta \in \mathbb{R}^d$ - входной параметр

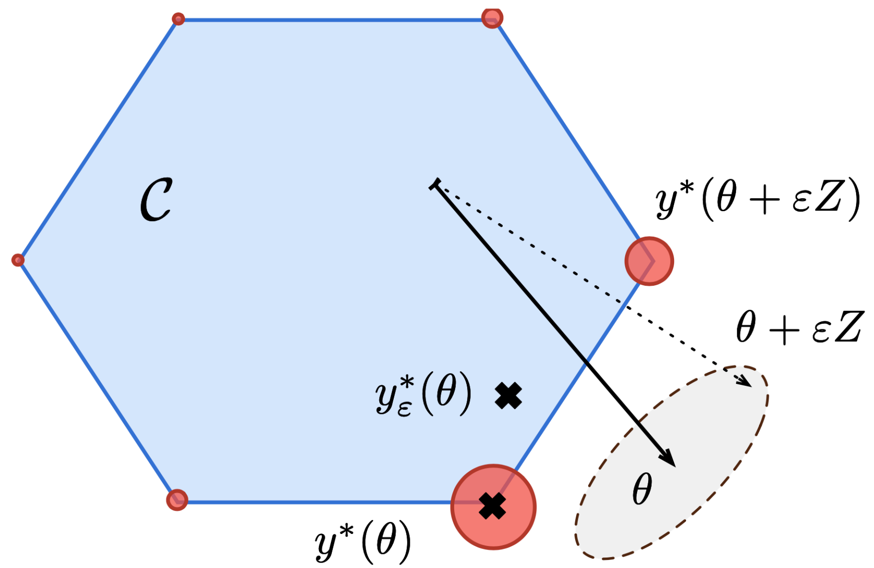


$$F(\theta) = \max_{y \in \mathcal{C}} \langle y, \theta \rangle \qquad y^*(\theta) = \arg\max_{y \in \mathcal{C}} \langle y, \theta \rangle$$

$$y^*(\theta) = \nabla_\theta F(\theta)$$

Возмущенный максимизатор (Perturbed maximizer)

Добавляем шум:



εZ - вектор шума

$\varepsilon > 0$ - температура

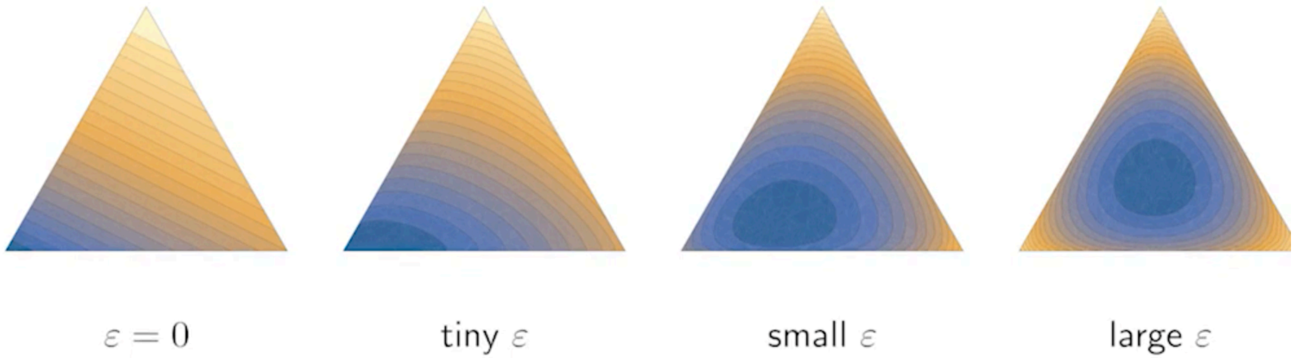
Z имеет положительную дифференцируемую плотность

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

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

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

Свойства



Связь с регуляризацией

$\varepsilon\,\Omega = (F_\varepsilon)^\star$ выпуклая функция с областью определения C :

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

Поведение при экстремальных температурах

При $\varepsilon \to 0$: $y_\varepsilon^*(\theta) \to y^*(\theta)$ **для уникального max**

При $\varepsilon \to \infty$: $y_\varepsilon^*(\theta) \to \arg\min_{y\in\mathcal{C}} \Omega(y)$

Свойства

Дифференцируемость

Для шума Z с распределением $\mathrm{d}\mu(z) \propto \exp(-\nu(z))\mathrm{d}z$ и дважды дифференцируемой ν, справедливо следующее утверждения:

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

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

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

Практическая реализация

$y^*_\varepsilon(\theta) = \underset{y \in \mathcal{C}}{\arg\max} \left\{ \langle y, \theta \rangle - \varepsilon\, \Omega(y) \right\}$ - хотим оценку на возмущенный максимизатор и его Якобиан

Нужно: $(Z^{(1)}, \ldots, Z^{(M)})$ - независимо одинаково распределённые с.в., $Z \sim \mu$

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

$\mathbf{E}[y^{(m)}] = y^*_\varepsilon(\theta)$ для любого $m \in \{1, \ldots, M\}$

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

Обучение: функция потерь

Функция потерь Фенхеля-Янга: $L_\varepsilon(\theta\,;y) = F_\varepsilon(\theta) + \varepsilon\,\Omega(y) - \langle\theta, y\rangle$

- Неотрицательна
- Выпукла
- Минимум равный 0 тогда и только тогда, когда θ: $y_\varepsilon^*(\theta) = y$

Градиент функции: $\nabla_\theta L_\varepsilon(\theta\,;y) = \nabla_\theta F_\varepsilon(\theta) - y = y_\varepsilon^*(\theta) - y$

Обучение

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

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

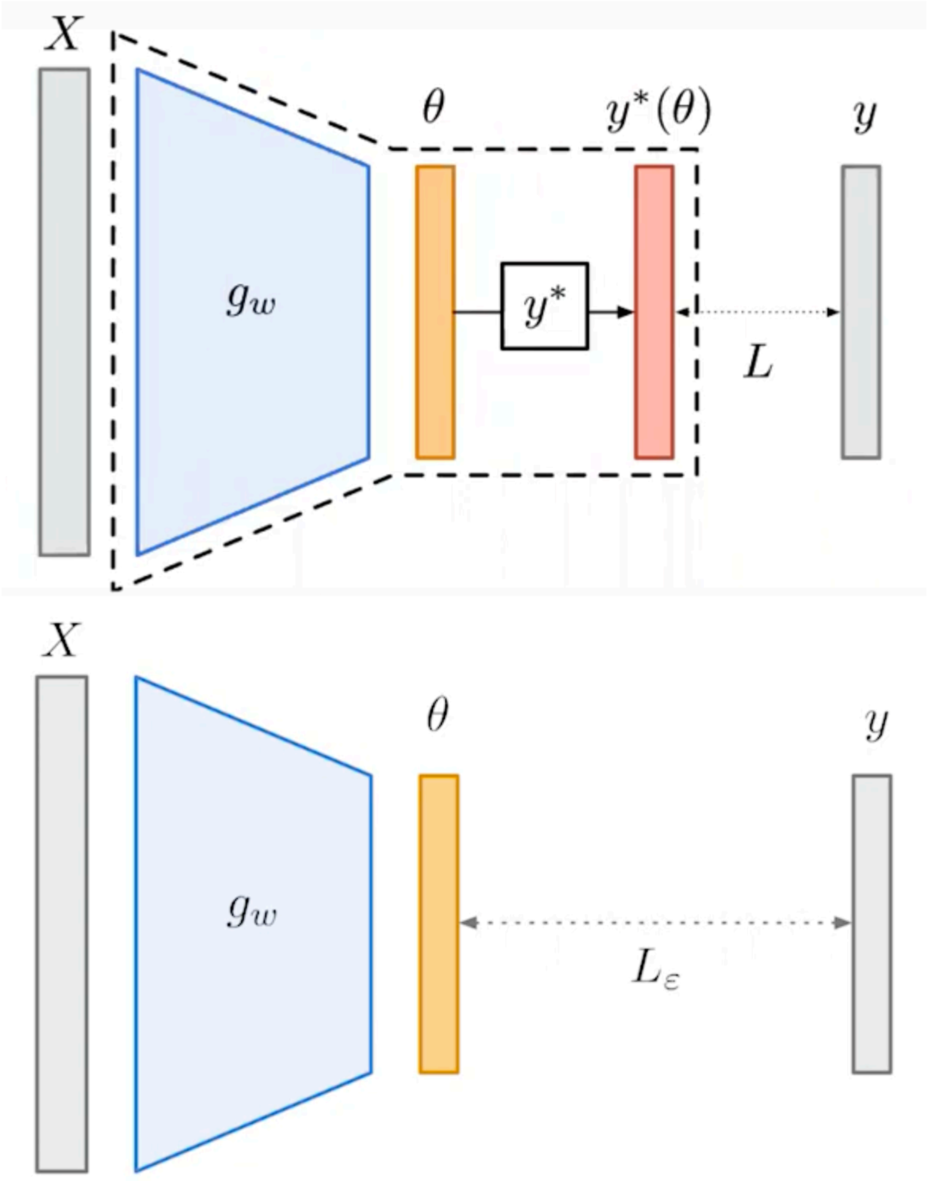
+

Дважды стохастичный градиент (doubly stochastic scheme) $\nabla_w L_\varepsilon(g_w(x_i)\,;y_i)$

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

Выводы

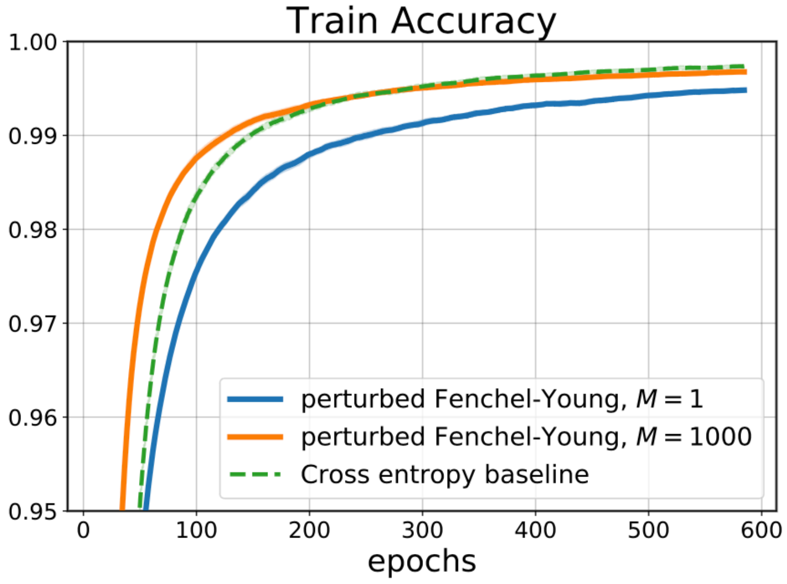
- Предложен новый универсальный метод преобразования дискретных оптимизаторов, который применим к любому солверу black-box модели.
- Метод позволяет дифференцировать argmax через возмущенный максимизатор.
- Полученные производные выражаются в виде простых мат ожиданий, которые легко аппроксимируются методами Монте-Карло.
- Произведена интеграция с функцией потерь Фенхеля-Янга, для удобного применения

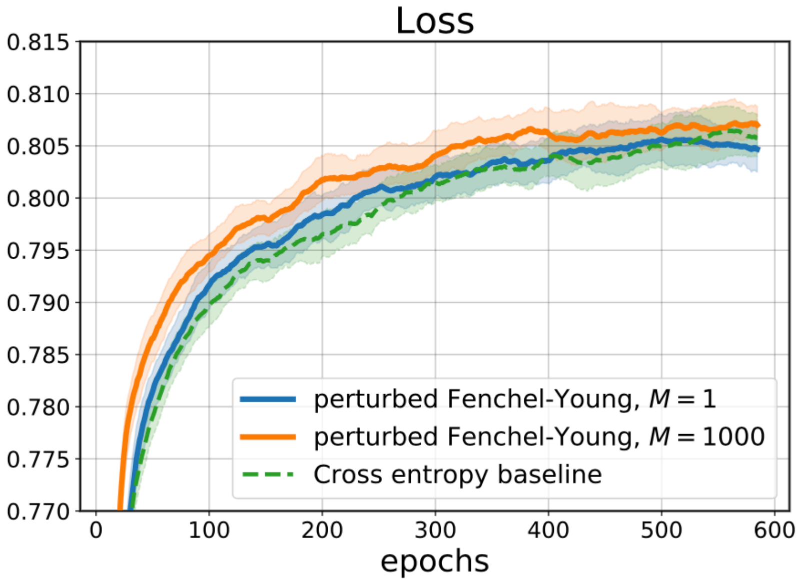


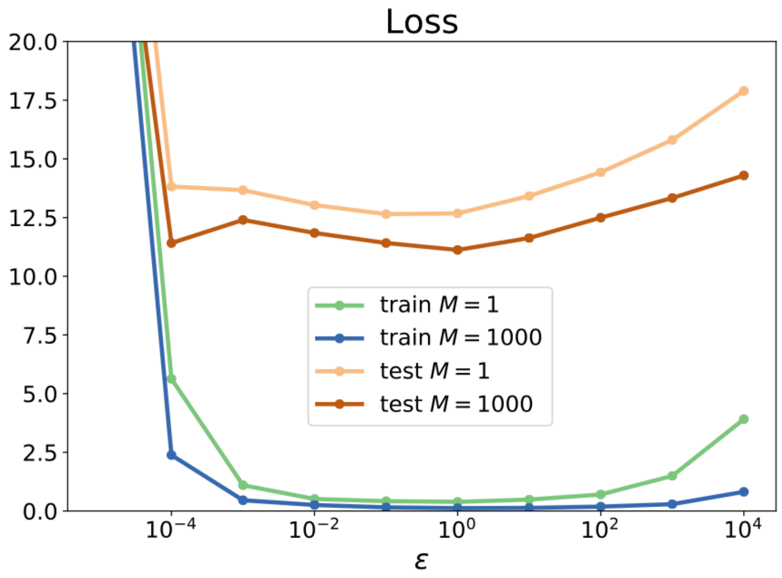
Эксперименты

Классификация

- возмущенный максимизатор с гауссовским шумом
- CIFAR-10
- vanilla-CNN (4 сверточных слоя and 2 полносвязных слоя), 600 эпох, размер батча = 32



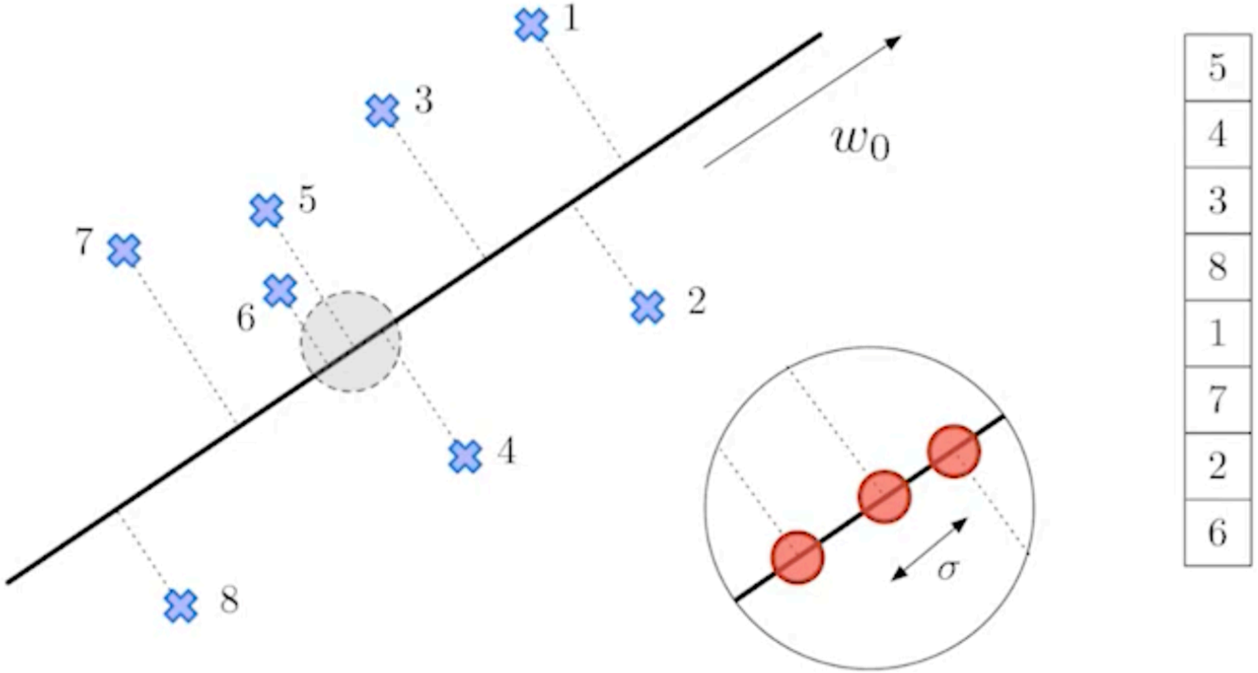




Эксперименты

Ранжирование

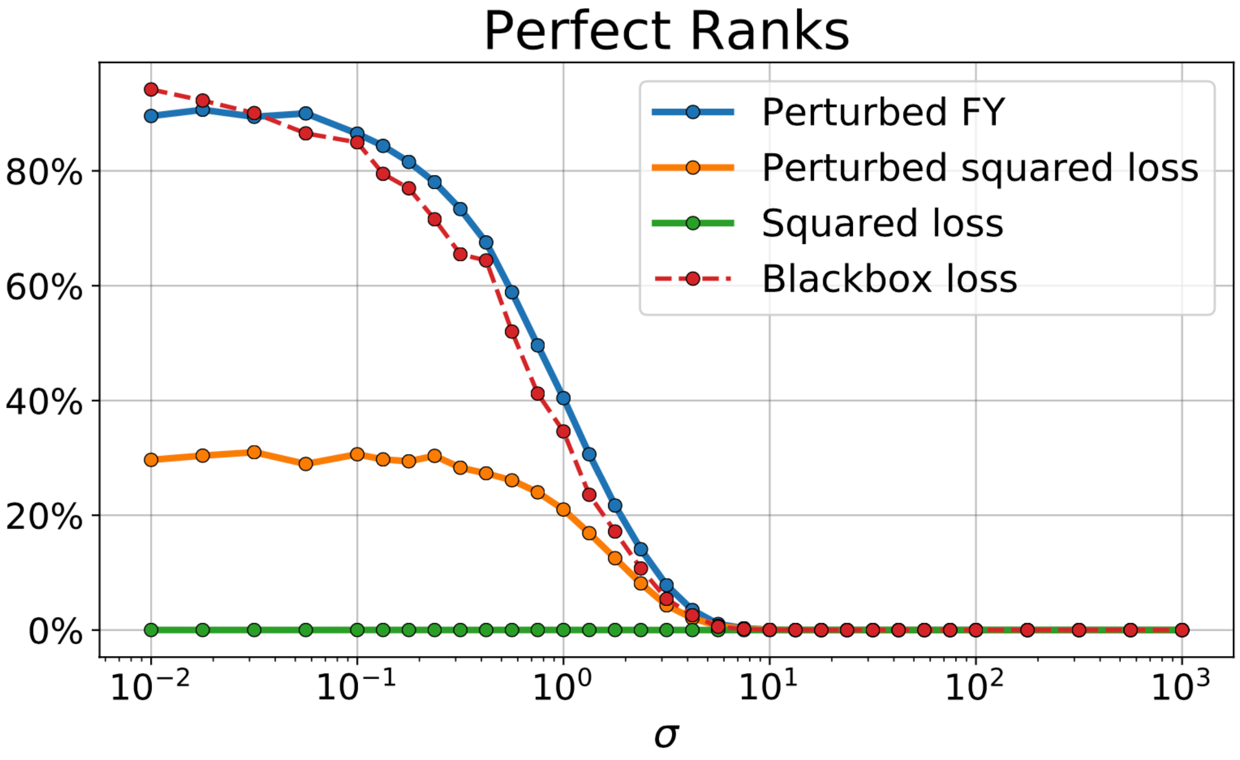
$$y_i = \arg\max_y \langle x_i^\top w_0 + \sigma Z_i, y \rangle$$



Эксперименты

Ранжирование

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



Perturbed Fenchel-Young (proposed)

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

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

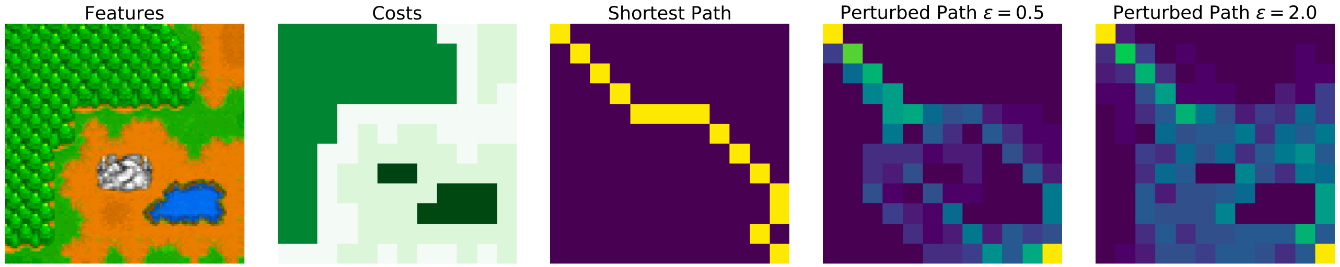
Blackbox loss: $\frac{1}{2}\|y_i - y^*(g_w(x_i))\|^2$

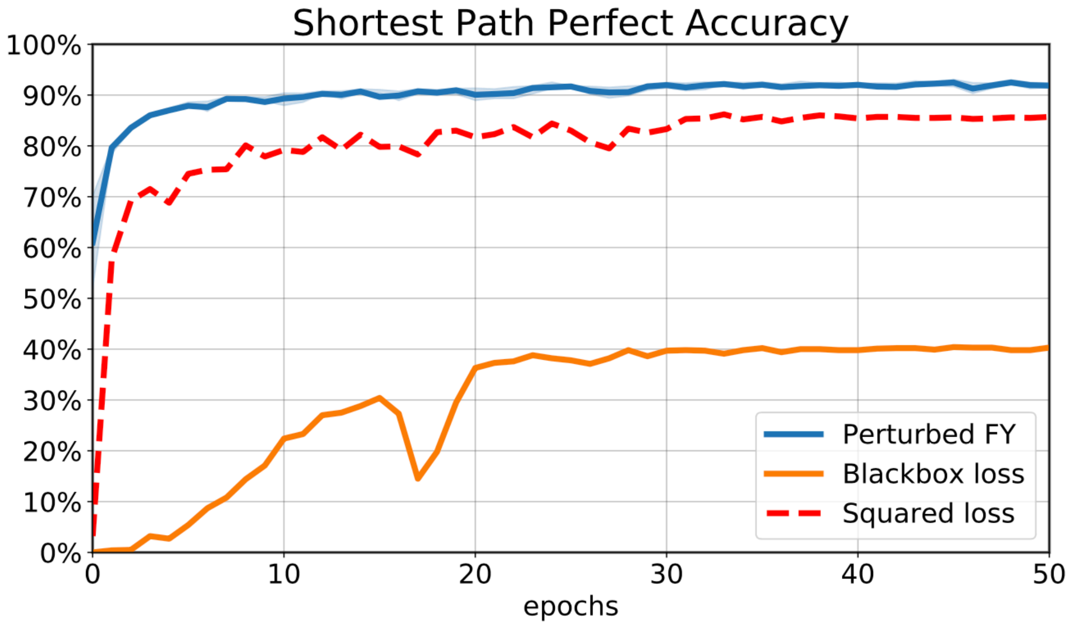
Используется приближение градиента из статьи Vlastelica M. Differentiation of blackbox combinatorial solvers, 2019.

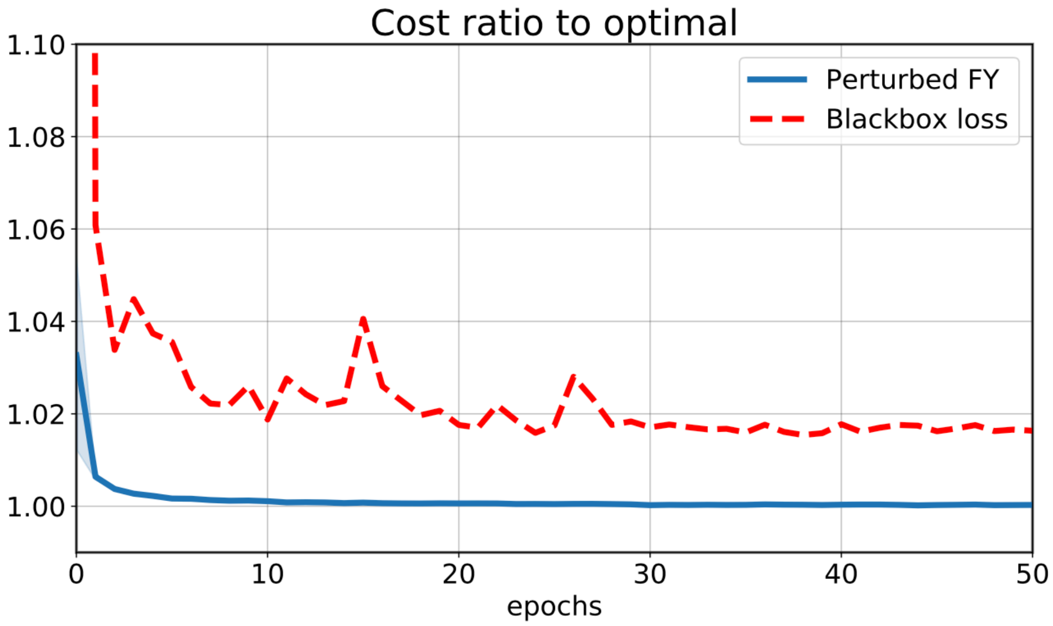
Эксперименты

Поиск кратчайшего пути

- Первые 5 слоев ResNet18
- 50 эпох, размер батча = 70
- ε = 1 , M = 1







Вопросы

1. Зачем авторы статьи используют возмущенный максимизатор, какие задачи решаются при использовании?
2. Выпишите формулу функции потерь Фенхеля-Янга и объясните все ее обозначения.
3. Опишите любой из 3 экспериментов и выводы, к которым они приводят.

Источники

- Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, Francis Bach. Learning with Differentiable Perturbed Optimizers, 2020.