Large-scale Stochastic Optimization of NDCG Surrogates for Deep Learning with Provable Convergence

Zi-Hao Qiu*[1] Quanqi Hu*[2] Yongjian Zhong[2]
Lijun Zhang[1] Tianbao Yang[2]

[1]Nanjing University [2]the University of Iowa

*Equal Contribution

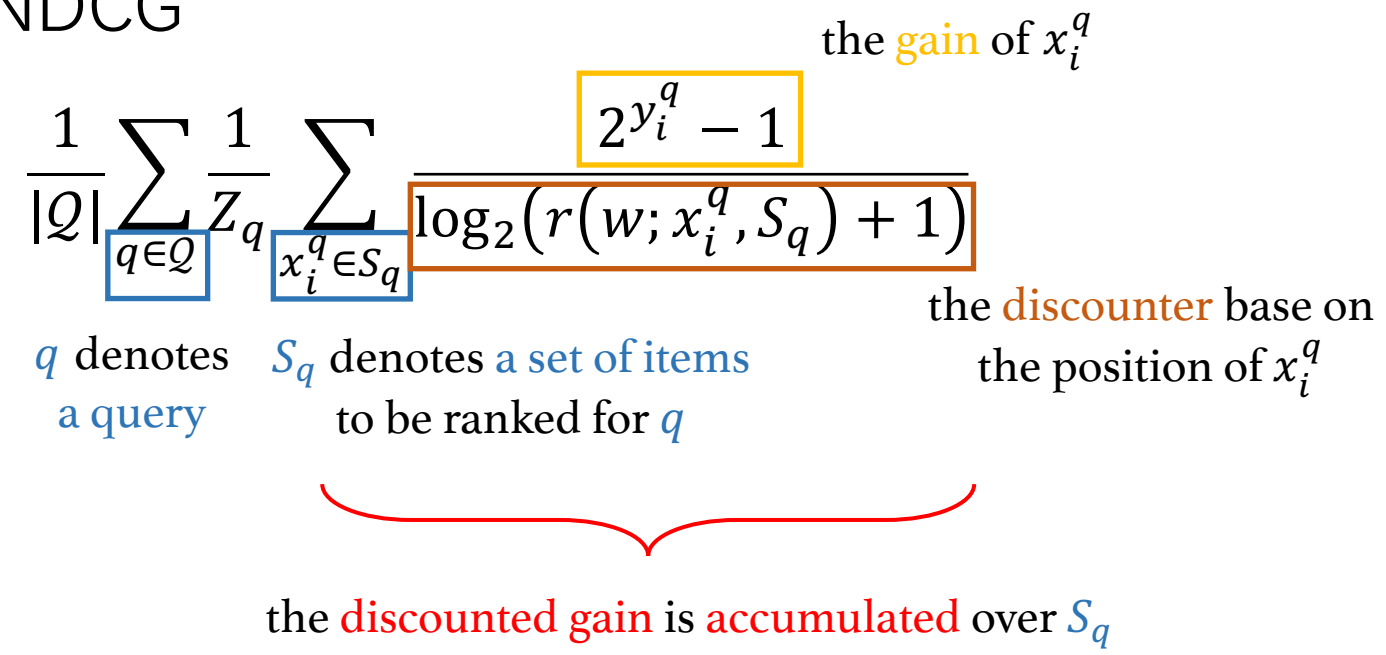
Introduction

- Normalized Discounted Cumulative Gain (NDCG)
  - a famous measure of ranking quality
  - widely used in recommender systems, learning to rank, …

Introduction

- Normalized Discounted Cumulative Gain (NDCG)
  - a famous measure of ranking quality
  - widely used in recommender systems, learning to rank, …

- Definition of NDCG



Introduction

- Normalized Discounted Cumulative Gain (NDCG)
  - a famous measure of ranking quality
  - widely used in recommender systems, learning to rank, …

- Definition of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)}$$

- The top-K variant of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\mathbb{I}\left(x_i^q\in S_q[K]\right)\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)}$$

sum over items in the top K positions in the ordered list

Introduction

- Normalized Discounted Cumulative Gain (NDCG)
  - a famous measure of ranking quality
  - widely used in recommender systems, learning to rank, ⋯
- Definition of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r(w;x_i^q,S_q)+1\right)}$$

Efficient and provable stochastic methods for optimizing NDCG are lacking.

- The top-K variant of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\mathbb{I}\left(x_i^q\in S_q[K]\right)\frac{2^{y_i^q}-1}{\log_2\left(r(w;x_i^q,S_q)+1\right)}$$

sum over items in the top K positions in the ordered list

NDCG Optimization

- Formulation of the optimization of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S} f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

NDCG Optimization

- Formulation of the optimization of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \Longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S} f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

Finite-sum
Coupled
Compositional
Optimization

NDCG Optimization

- Formulation of the optimization of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S} f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

Finite-sum Coupled Compositional Optimization

- Formulation of the optimization of top-K NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\mathbb{I}\left(x_i^q\in S_q[K]\right)\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S}\psi\left(h_q\left(x_i^q;w\right)-\hat{\lambda}_q(w)>0\right)f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

$$\text{s.t. } \hat{\lambda}_q(w)=\arg\min_{\lambda} L(\lambda,w;K,S_q),\ \forall q\in\mathcal{Q}$$

NDCG Optimization

- Formulation of the optimization of NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S} f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

Finite-sum Coupled Compositional Optimization

- Formulation of the optimization of top-K NDCG

$$\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{Z_q}\sum_{x_i^q\in S_q}\mathbb{I}\left(x_i^q\in S_q[K]\right)\frac{2^{y_i^q}-1}{\log_2\left(r\left(w;x_i^q,S_q\right)+1\right)} \longrightarrow \min_w \frac{1}{|S|}\sum_{(q,x_i^q)\in S}\psi\left(h_q\left(x_i^q;w\right)-\hat{\lambda}_q(w)>0\right)f_{q,i}\left(g\left(w;x_i^q,S_q\right)\right)$$

$$\text{s.t. } \hat{\lambda}_q(w)=\arg\min_{\lambda} L(\lambda,w;K,S_q),\forall q\in\mathcal{Q}$$

Multi-Block Bilevel Optimization

NDCG Optimization

- Key challenge

$$f_{q,i}\left(g\left(w; x_i^q, S_q\right)\right) = \frac{2^{y_i^q} - 1}{\log_2\left(g\left(w; x_i^q, S_q\right) + 1\right)}$$

$$g\left(w; x_i^q, S_q\right) = \sum_{x' \in S_q} l\left(h_q(x'; w) - h_q\left(x_i^q; w\right) \geq 0\right)$$

NDCG Optimization

- Key challenge

$$f_{q,i}\left(g\left(w; x_i^q, S_q\right)\right) = \frac{2^{y_i^q} - 1}{\log_2\left(g\left(w; x_i^q, S_q\right) + 1\right)}$$

$$g\left(w; x_i^q, S_q\right) = \sum_{x' \in S_q} l\left(h_q(x'; w) - h_q\left(x_i^q; w\right) \geq 0\right)$$

  - the inner function $g\left(w; x_i^q, S_q\right)$ involves $|S_q|$ items, which can be very large

NDCG Optimization

- Key challenge

$$f_{q,i}\left(g\left(w; x_i^q, S_q\right)\right) = \frac{2^{y_i^q} - 1}{\log_2\left(g\left(w; x_i^q, S_q\right) + 1\right)}$$

$$g\left(w; x_i^q, S_q\right) = \sum_{x' \in S_q} l\left(h_q(x'; w) - h_q\left(x_i^q; w\right) \geq 0\right)$$

  - the inner function $g\left(w; x_i^q, S_q\right)$ involves $|S_q|$ items, which can be very large
  - the outer function $f_{q,i}$ is non-linear, thus an unbiased stochastic gradient is not readily computed

NDCG Optimization

- Our strategy
  - use a *moving average estimator* to keep track of $g\left(w; x_i^q, S_q\right)$ for each x_i^q
  - $u_{q,i}^{(t+1)} = \gamma_0 g\left(w; x_i^q, B_q\right) + (1-\gamma_0) u_{q,i}^{(t)}$

$$\nabla_w f_{q,i}\left(g\left(w; x_i^q, B_q\right)\right) \Longrightarrow \nabla_w f_{q,i}\left(u_{q,i}\right) \nabla_w g\left(w; x_i^q, B_q\right)$$

NDCG Optimization

- Our strategy
  - use a *moving average estimator* to keep track of $g\left(w; x_i^q, S_q\right)$ for each x_i^q
  - $u_{q,i}^{(t+1)} = \gamma_0 g\left(w; x_i^q, B_q\right) + (1-\gamma_0) u_{q,i}^{(t)}$

$$\nabla_w f_{q,i}\left(g\left(w; x_i^q, B_q\right)\right) \Longrightarrow \nabla_w f_{q,i}\left(u_{q,i}\right) \nabla_w g\left(w; x_i^q, B_q\right)$$

  - intuitively, when t increases, w_{t-1} is getting close to w_t, hence the previous value of $u_{q,i}^{(t+1)}$ is helpful for estimating $\nabla_w f_{q,i}\left(g\left(w; x_i^q, S_q\right)\right)$
  - similar methods can be applied when optimizing the top-K NDCG variant

Algorithm

Algorithm 1 Stochastic Optimization of NDCG: SONG

Require: $\eta, \gamma_0, \beta_1, u^{(1)} = 0$
Ensure: $\mathbf{w}_T$
for $t = 1, \ldots T$ **do**
  Draw some relevant Q-I pairs $\mathcal{B} = \{(q, \mathbf{x}_i^q)\} \subset \mathcal{S}$
  For each sampled q draw a batch of items $\mathcal{B}_q \subset \mathcal{S}_q$
  for each sampled Q-I pair $(q, \mathbf{x}_i^q) \in \mathcal{B}$ **do**
    Let $\hat{g}_{q,i}(\mathbf{w}_t) = \frac{1}{|\mathcal{B}_q|} \sum_{\mathbf{x}' \in \mathcal{B}_q} \ell(\mathbf{w}_t; \mathbf{x}', \mathbf{x}_i^q, q)$
    Compute $u_{q,i}^{(t+1)} = (1 - \gamma_0) u_{q,i}^{(t)} + \gamma_0 \hat{g}_{q,i}(\mathbf{w}_t)$
    Compute $p_{q,i} = \nabla f_{q,i}(u_{q,i}^{(t)})$
  end for
  Compute the stochastic gradient estimator $G(\mathbf{w}_t)$ by

$$G(\mathbf{w}_t) = \frac{1}{|\mathcal{B}|} \sum_{(q, \mathbf{x}_i^q) \in \mathcal{B}} p_{q,i} \nabla \hat{g}_{q,i}(\mathbf{w}_t)$$

  Compute $\mathbf{m}_{t+1} = \beta_1 \mathbf{m}_t + (1 - \beta_1) G(\mathbf{w}_t)$
  update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{m}_{t+1}$
end for

After $T = O(1/\varepsilon^4)$ iterations, we can find an ε-stationary solution of $F(w)$, i.e., $\mathbb{E}[\|\nabla F(w_\tau)\|^2] \leq \varepsilon^2$ for a randomly selected $\tau \in \{1, \ldots, T\}$

Algorithm

Algorithm 2 Stochastic Optimization of top-K NDCG: K-SONG

Require: $\eta_0, \eta_1, \gamma_0, \gamma_0', \beta_1, u^{(1)} = 0, \lambda = 0$
Ensure: $\mathbf{w}_T$
for $t = 1, \ldots T$ **do**
  Draw some relevant Q-I pairs $\mathcal{B} = \{(q, \mathbf{x}_i^q)\} \subset \mathcal{S}$
  For each sampled q draw a batch of items $\mathcal{B}_q \subset \mathcal{S}_q$
  for each sampled Q-I pair $(q, \mathbf{x}_i^q) \in \mathcal{B}$ **do**
    Let $\hat{g}_{q,i}(\mathbf{w}_t) = \frac{1}{|\mathcal{B}_q|} \sum_{\mathbf{x}' \in \mathcal{B}_q} \ell(\mathbf{w}_t; \mathbf{x}', \mathbf{x}_i^q, q)$
    Let $u_{q,i}^{(t+1)} = (1 - \gamma_0) u_{q,i}^{(t)} + \gamma_0 \hat{g}_{q,i}(\mathbf{w}_t)$
    Let $p_{q,i} = \psi(h_q(\mathbf{x}_i^q; \mathbf{w}_t) - \lambda_{q,t}) \nabla f_{q,i}(u_{q,i}^t)$
  end for
  for each sampled query $q \in \mathcal{B}$ **do**
    Let $s_{q,t+1} = (1 - \gamma_0') s_{q,t} + \gamma_0' \nabla_\lambda^2 L_q(\lambda_{q,t}; \mathbf{w}_t; \mathcal{B}_q)$
    Let $\lambda_{q,t+1} = \lambda_{q,t} - \eta_0 \nabla_\lambda L_q(\lambda_{q,t}; \mathbf{w}_t; \mathcal{B}_q)$
  end for
  Compute a stochastic gradient $G(\mathbf{w}_t)$ according to (5) or (6)
  Compute $\mathbf{m}_{t+1} = \beta_1 \mathbf{m}_t + (1 - \beta_1) G(\mathbf{w}_t)$
  Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_1 \mathbf{m}_{t+1}$
end for

After $T = O(1/\varepsilon^4)$ iterations, we can find an ε-stationary solution of $F(w)$, i.e., $\mathbb{E}[\|\nabla F(w_\tau)\|^2] \leq \varepsilon^2$ for a randomly selected $\tau \in \{1, \ldots, T\}$

Experiments

Table 1: The test NDCG on four datasets. We report the average NDCG@3 for two LTR datasets, the average NDCG@20 for two RS datasets, and standard deviation (within brackets) over 3 runs with different random seeds. Full results are in Appendix D.3.

METHOD	NDCG@3		NDCG@20	
	MSLE WEB30K	YAHOO! LTR	MOVIELENS20M	NETFLIX PRIZE
RANKNET	0.5105±0.0004	0.7150±0.0004	0.0744±0.0013	0.0489±0.0003
LISTNET	0.5058±0.0001	0.7151±0.0004	0.0875±0.0004	0.0700±0.0002
LISTMLE	0.5074±0.0002	0.7146±0.0006	0.0799±0.0001	0.0508±0.0004
LAMBDARANK	0.5118±0.0003	0.7155±0.0002	0.0913±0.0002	0.0693±0.0002
APPROXNDCG	0.5114±0.0005	0.7152±0.0007	0.0938±0.0003	0.0592±0.0009
NEURALNDCG	0.5101±0.0005	0.7139±0.0001	0.0901±0.0003	0.0718±0.0003
SONG	0.5136±0.0006	0.7187±0.0004	0.0969±0.0002	**0.0749**±0.0002
K-SONG	**0.5147**±0.0006	**0.7191**±0.0004	**0.0973**±0.0003	0.0743±0.0003

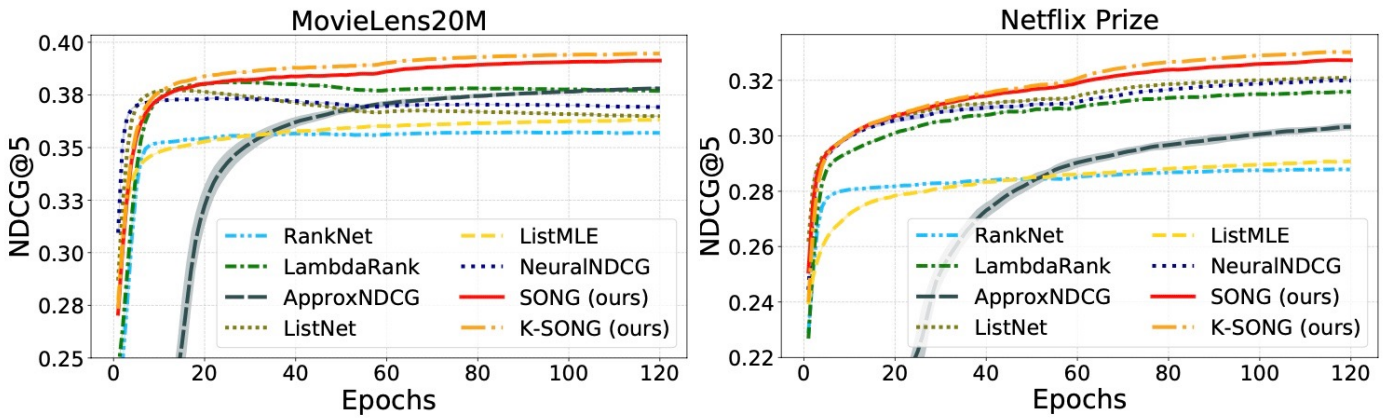
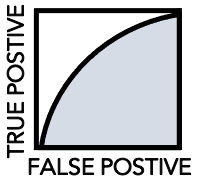


Figure 1: Comparison of convergence of different methods in terms of validation NDCG@5 scores on two RS datasets.

Thank You

Paper: https://arxiv.org/abs/2202.12183
Project website: https://libauc.org



LibAUC