Macchine di Boltzmann per il Topic Modeling

Il modello Replicated Softmax nella libreria OCTIS

Relatore:
Prof. Antonio Candelieri
Correlatore:
Prof. Elisabetta Fersini

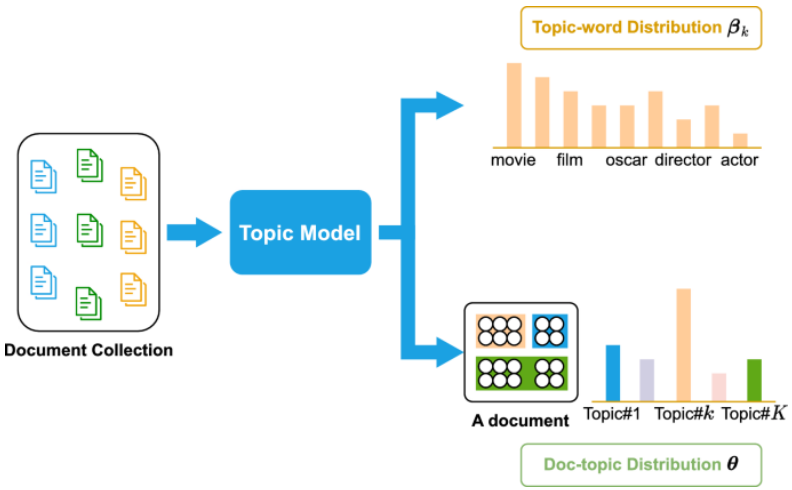
Candidato:
Federico Rausa

Università degli Studi Milano-Bicocca

21 Gennaio 2026

Introduzione

- **Topic Modeling**: Tecnica non supervisionata per l'analisi di una collezione di documenti. Utile per costruire i topic, variabili latenti interpretabili come distribuzioni di parole, e rappresentare il testo come una mistura di topic.
- **Replicated Softmax** (Hinton et al. 2009): Topic Model basato sull'utilizzo di una Macchina di Boltzmann Ristretta.
- **OCTIS** [Optimizing and Comparing Topic Models is Simple](Fersini, Candelieri et al. 2021): Python Package pensato per la gestione integrata di un'analisi di topic modeling. Include preprocessing, diversi topic models, metriche di validazione e strumenti di ottimizzazione bayesiana.



output di un topic model

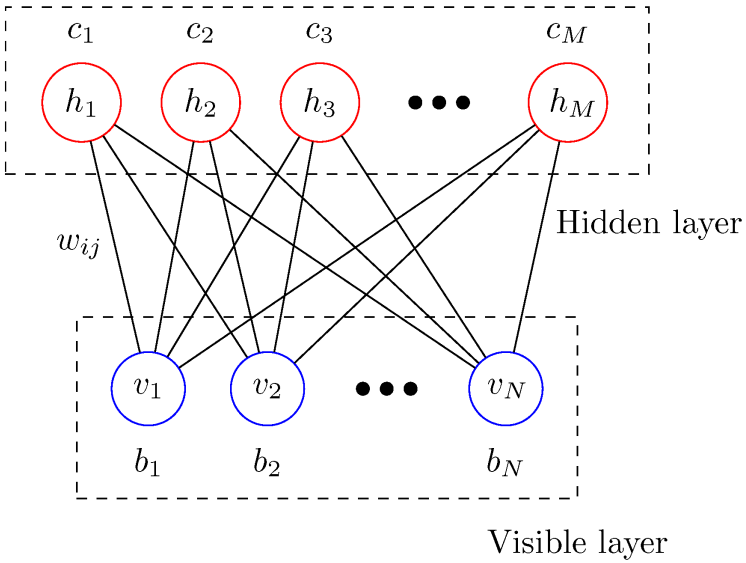
- La matrice topic-document (di probabilità o score) di dimensioni KxN
- La matrice topic-word (di probabilità o score) di dimensioni KxV
- La matrice di descrizione dei topic (di stringhe/parole/token) di dimensioni KxG,

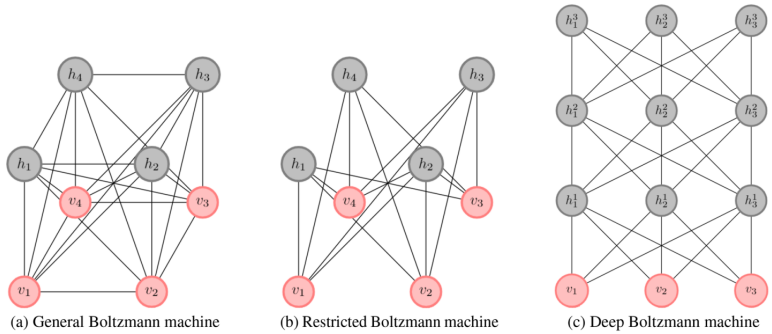
con K = # topic, N = # documenti = |corpus|,
V = # parole = |vocabolario| , G arbitrario (di solito 10)

metriche di validazione per il topic modeling:

- **entropia**: perplexity, entropia di Shannon. Valutazioni a livello di Topic-Document e Topic-Word.
- **Metriche di classificazione**: Accuracy, F1-score, Precision, Recall (richiedono un etichettamento a priori e un modello supervisionato). Valutazioni a livello di Topic-Document.
- **metriche di Coherence/Coerenza**: Cv, UMass , UCI, NPMI. Valutazioni a livello di Topic-description.
- **Topic Diversity** (Dieng et al. 2020). Valutazioni a livello di Topic-description.

Macchine Ristrette di Boltzmann




(a) General Boltzmann machine (b) Restricted Boltzmann machine (c) Deep Boltzmann machine

funzione di energia del modello RS:

$$-E(\mathbf{v},\mathbf{h}) = \sum_{j=1}^{F}\sum_{k=1}^{K} v_k h_j w_{kj} + \sum_{k=1}^{K} v_k a_k + D\sum_{j=1}^{F} h_j b_j \qquad (1)$$

dove D è il numero delle parole nel documento (size della multinomiale).

Energia e distribuzione di Boltzmann

Free energy e funzione di partizione:

$$P(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h} \in \mathbb{H}} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{2}$$

$$-F(\mathbf{v}) = -\ln(p(\mathbf{v})) = \sum_{i=1}^{V} v_i a_i + \sum_{j=1}^{H} \ln\left(1 + \exp\left(b_j D + \sum_{i=1}^{V} v_i w_{ij}\right)\right) \tag{3}$$

$$Z = \sum_{\mathbf{v} \in \mathbb{V}} \sum_{\mathbf{h} \in \mathbb{H}} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{4}$$

Probabilità condizionate dei layer:

$$P(h = 1|\mathbf{v}) = \sigma\left(b_j + \sum_{k=1}^{K} w_{kj} v_k\right) \tag{5}$$

$$P(v_{ik} = 1|\mathbf{h}) = \frac{\exp\left(a_k + \sum_{j=1}^{F} w_{kj} h_j\right)}{\sum_{k'=1}^{K} \exp\left(a_{k'} + \sum_{j=1}^{F} w_{k'j} h_j\right)} \tag{6}$$

Gradient Descent e Contrastive Divergence

Sotto il modello, si dimostra l'equivalenza tra perplexity, log verosimiglianza, free energy e divergenza di Kullback-Leibler:

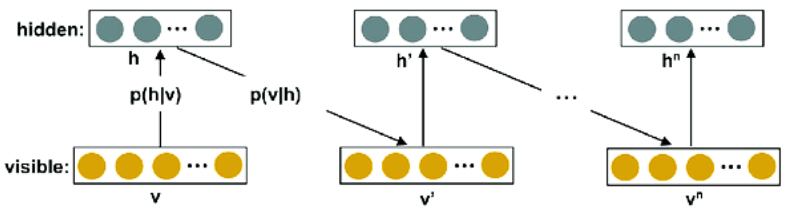
$$argmin_\theta \ \mathrm{KL}(q(x)||P(x;\theta)) = argmax_\theta \ln(P(x;\theta)) = argmin_\theta F(x;\theta) = argmin_\theta \mathrm{PPL}(x, P(x;\theta)) \tag{7}$$

Preso un batch di N documenti, i gradienti medi dei parametri sono i seguenti:

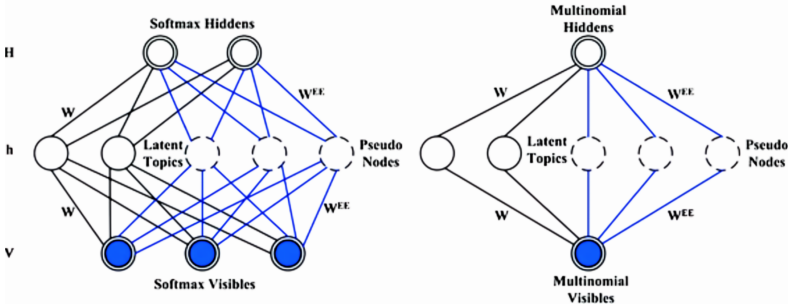
$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln(P(\mathbf{v}_n))}{\partial w_{kj}} = \mathbb{E}_{data}[v_k h_j] - \mathbb{E}_{model}[v_k h_j] \tag{8}$$

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln(P(\mathbf{v}_n))}{\partial b_{j}} = \mathbb{E}_{data}[h_j] - \mathbb{E}_{model}[h_j] \tag{9}$$

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln(P(\mathbf{v}_n))}{\partial a_{k}} = \mathbb{E}_{data}[v_k] - \mathbb{E}_{model}[v_k] \tag{10}$$



Over-Replicated Softmax



funzione di energia:

$$-E(\mathbf{v},\mathbf{h},\mathbf{H}) = \sum_{k=1}^{K}\sum_{j=1}^{F} w_{jk} h_j (v_k + g_k) + \sum_{k=1}^{K} a_k (v_k + g_k) + (M+D)\sum_{j=1}^{F} b_j h_j \tag{11}$$

$$P(h_j = 1|\mathbf{v},\mathbf{H}) = \sigma\left(b_j + \sum_{k=1}^{K} w_{jk}(v_k + g_k)\right) \tag{12}$$

$$P(H_{ik} = 1|\mathbf{h}) = P(v_{ik} = 1|\mathbf{h}) = \frac{\exp\left(a_k + \sum_{j=1}^{F} w_{kj} h_j\right)}{\sum_{k'=1}^{K} \exp\left(a_{k'} + \sum_{j=1}^{F} w_{k'j} h_j\right)} \tag{13}$$

Ottimizzazione del training

Tecniche di ottimizzazione adottate

- penalizzazione del gradiente (L1 o L2)
- utilizzo di Stocastic-Batch Gradient Descent
- utilizzo di ottimizzatori di GD (Adagrad, Momentum, RMSprop e Adam)
- scelta degli iperparametri con ottimizzazione bayesiana

Componenti di un processo di ottimizzazione bayesiana:

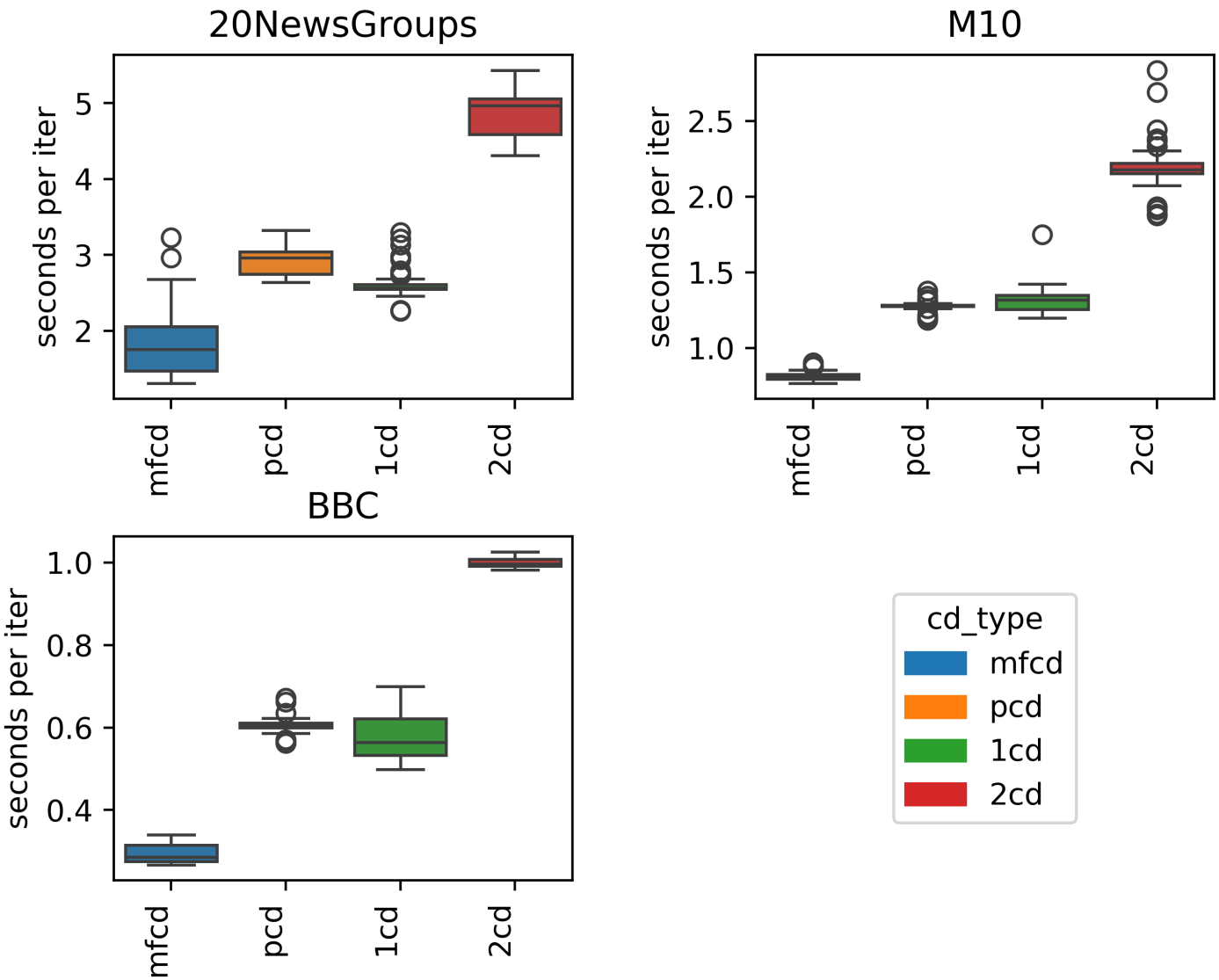
- funzione obiettivo
- spazio degli iperparametri
- modello surrogato (GP, RF ..)
- funzione di acquisizione (LCB, EI, PI ..)

Table: dataset di octis

Name in OCTIS	num Docs	num Words	num Labels	Language
20NewsGroup	16309	1612	20	English
BBC_News	2225	2949	5	English
M10	8355	1696	10	English

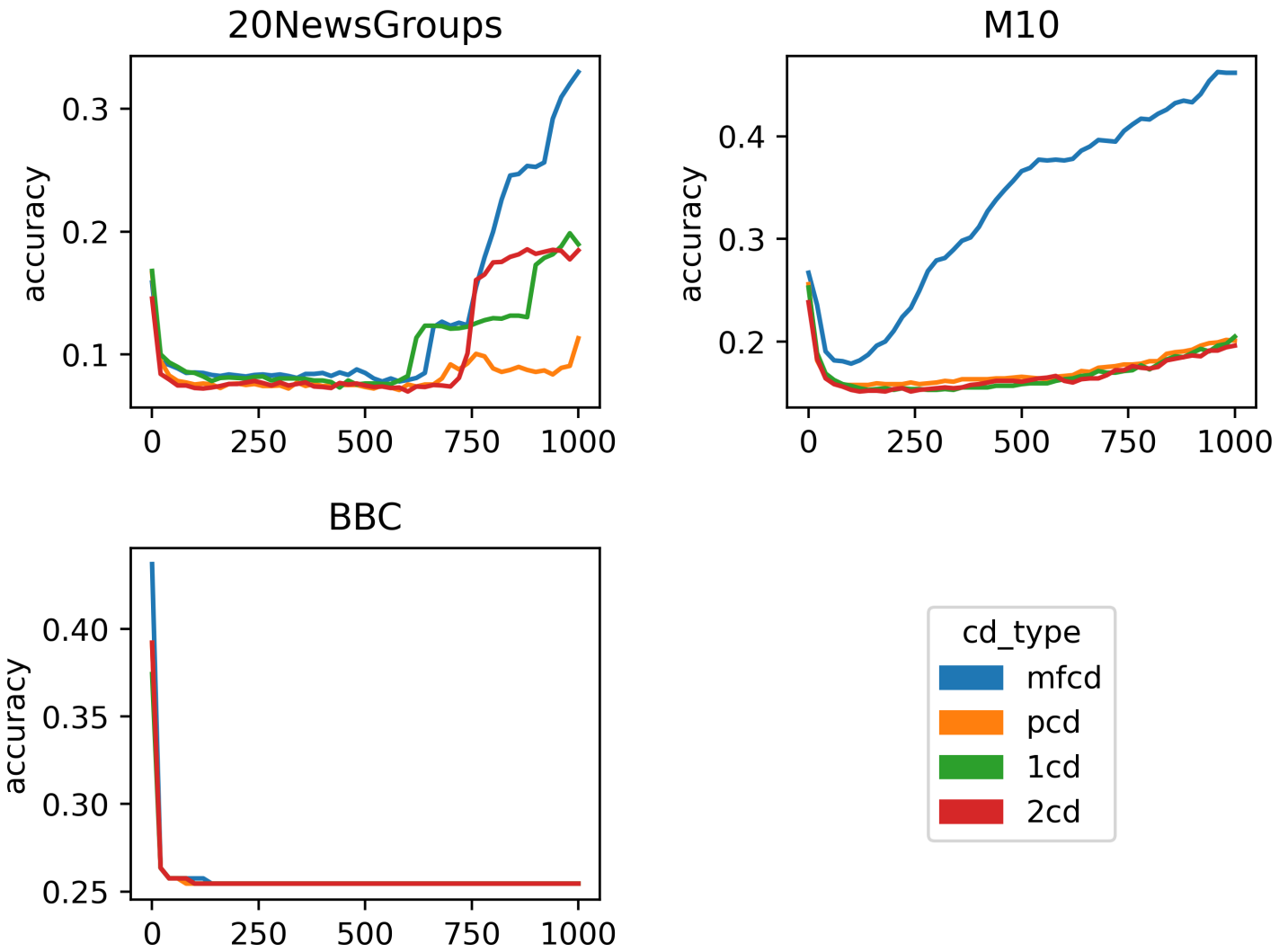
Confronto tra metodi di contrastive divergence

Secondi per iterazione con Replicated Softmax



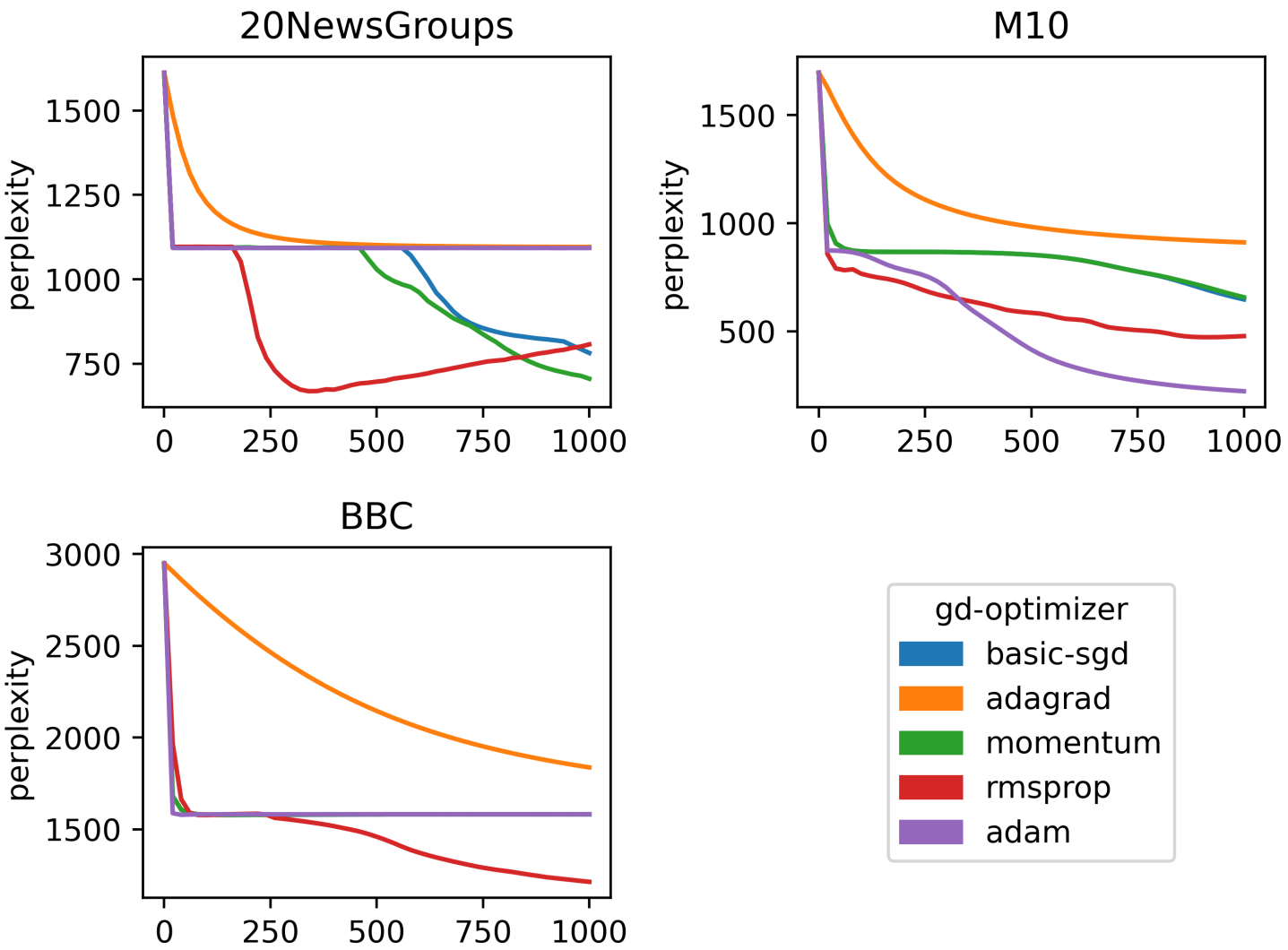
Confronto tra metodi di contrastive divergence

Accuracy per Replicated Softmax



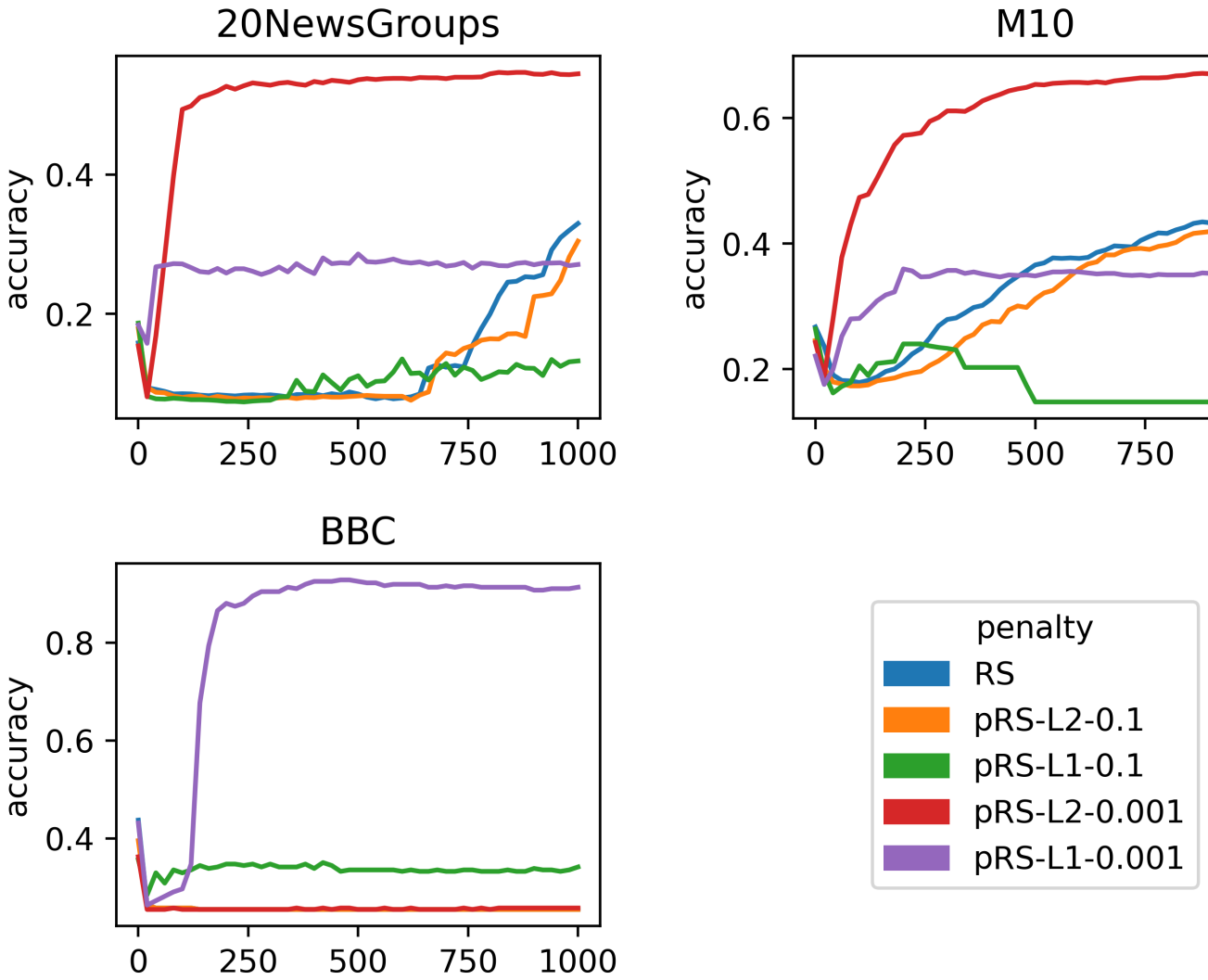
Ottimizzazione del gradiente

Perplexity per Replicated Softmax



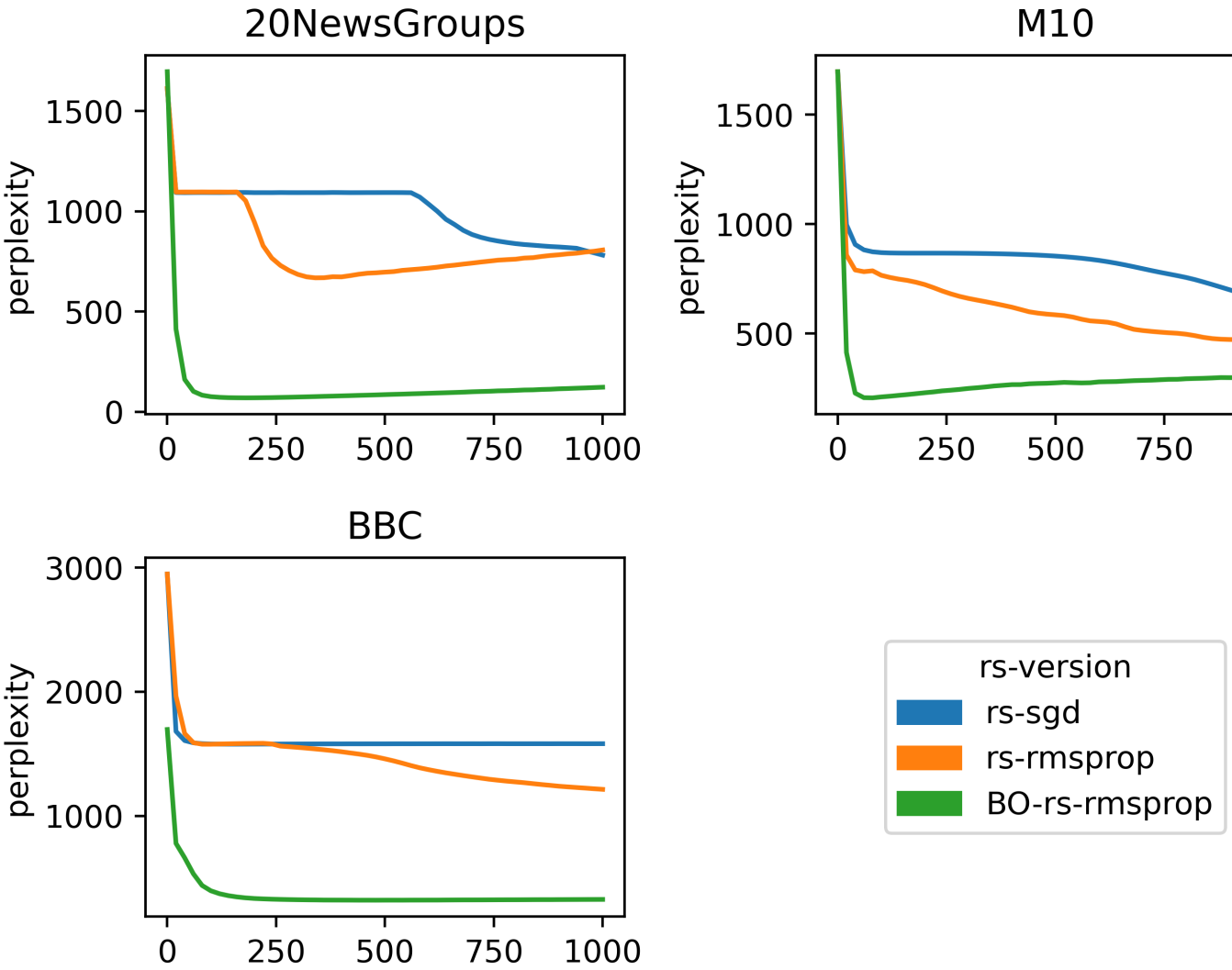
Penalizzazione del modello

Accuracy per penalized Replicated Softmax



Ottimizzazione Bayesiana

Ottimizzazione bayesiana su Topic Diversity con modello surrogato GP e acq. func. UCB:



Conclusioni

Risultati rilevanti:

- guadagno di efficienza importante con BO e MFCD
- Miglioramento di performance significativo con RMSprop e Adam
- Miglioramento dell'interpretabilità dei topic con penalized-RS e over-RS

Limiti del lavoro:

- ridotto numero di epoche
- problema della scelta del numero di topic
- confronto con VAE per il topic modeling

Grazie per l'attenzione

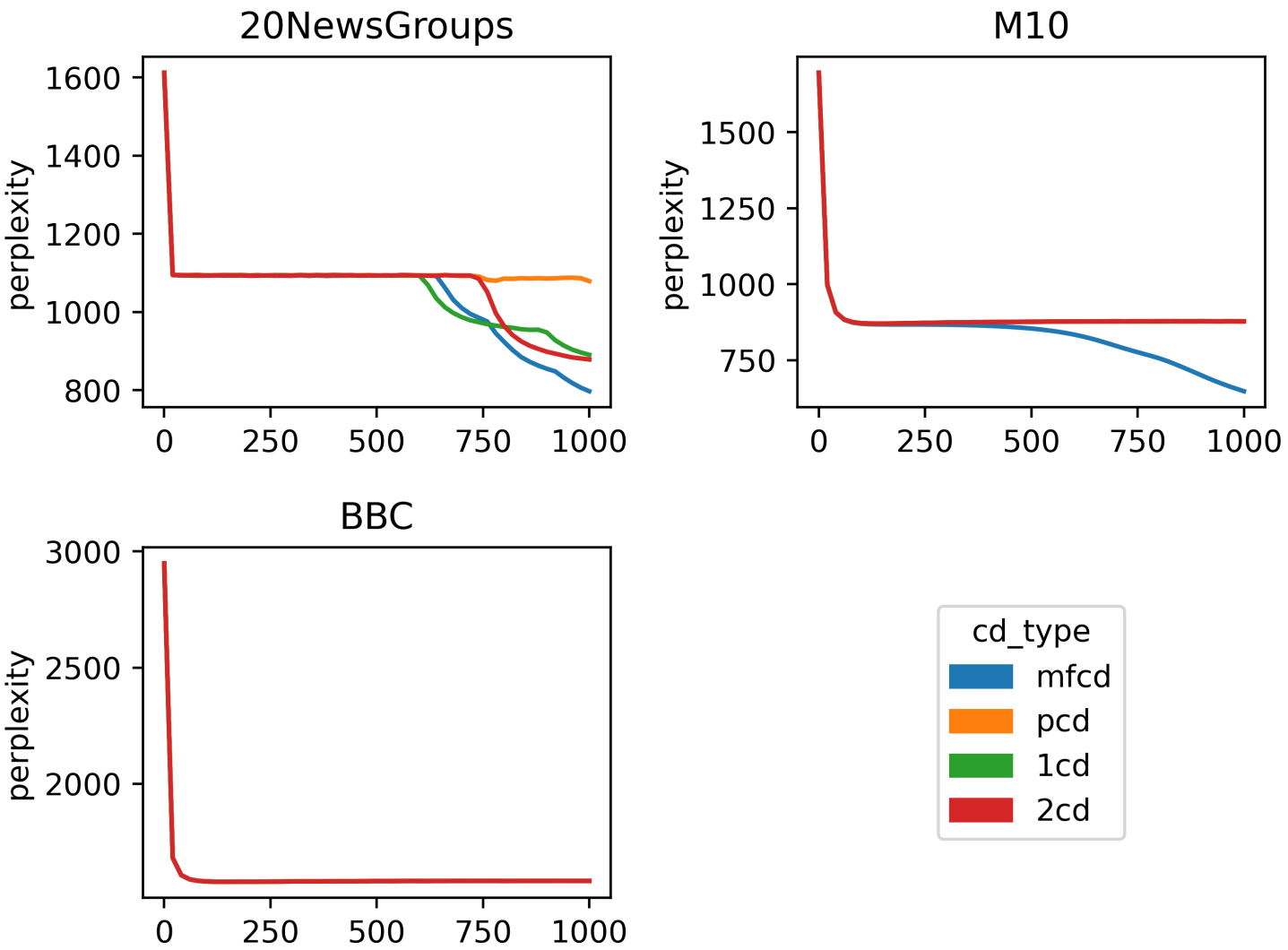
Interpretazione dei topic

BBC: Over-RS con ottimizzazione rmsprop

	$word_1$	$word_2$	$word_3$	$word_4$	$word_5$
$topic_3$	game	player	play	film	season
$topic_4$	election	tax	labour	economy	taxis
$topic_5$	firm	company	government	analyst	market
$topic_1$	include	give	work	back	add
$topic_2$	film	win	man	game	work

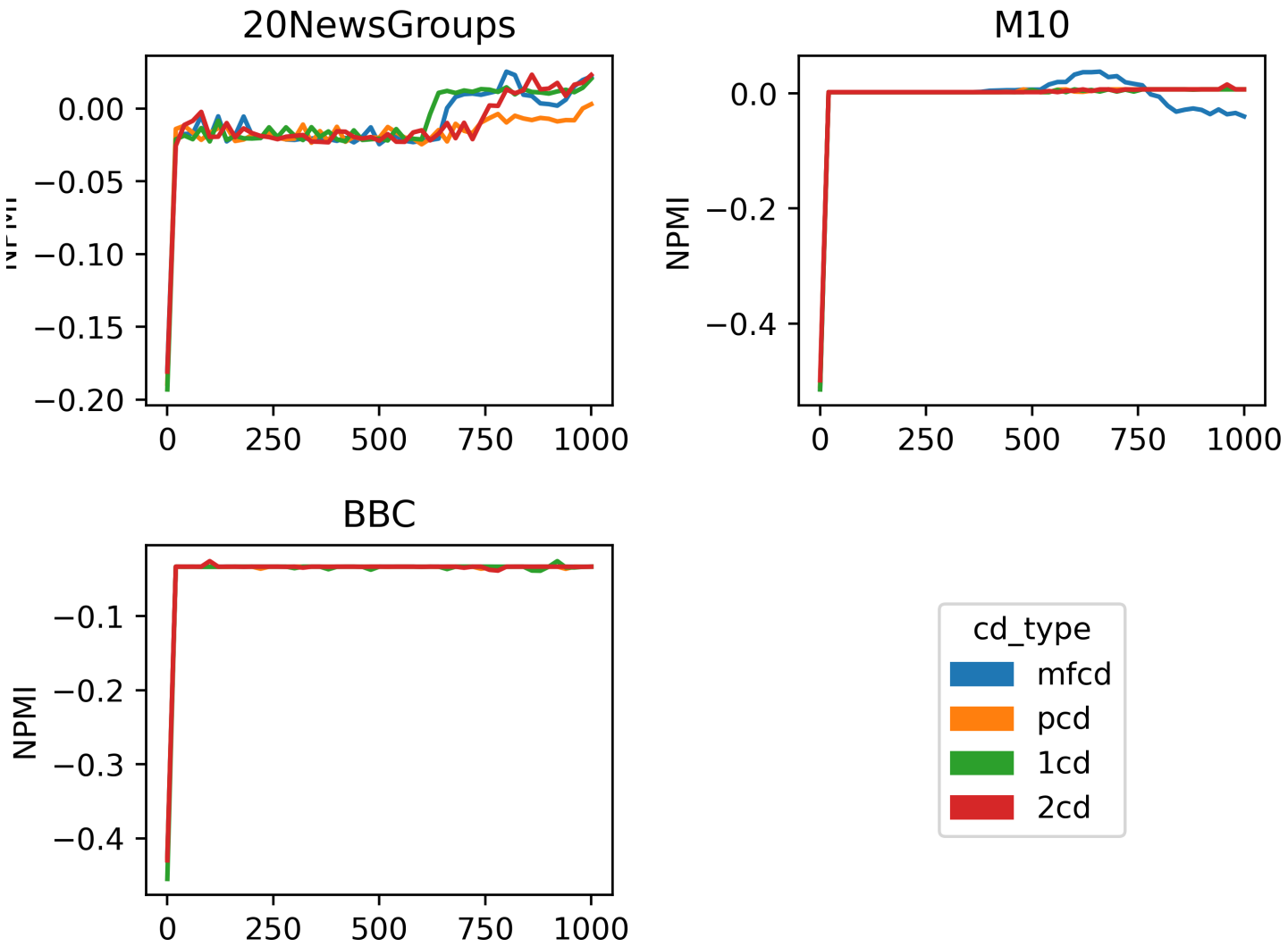
Confronto tra metodi di contrastive divergence

Perplexity per Replicated Softmax



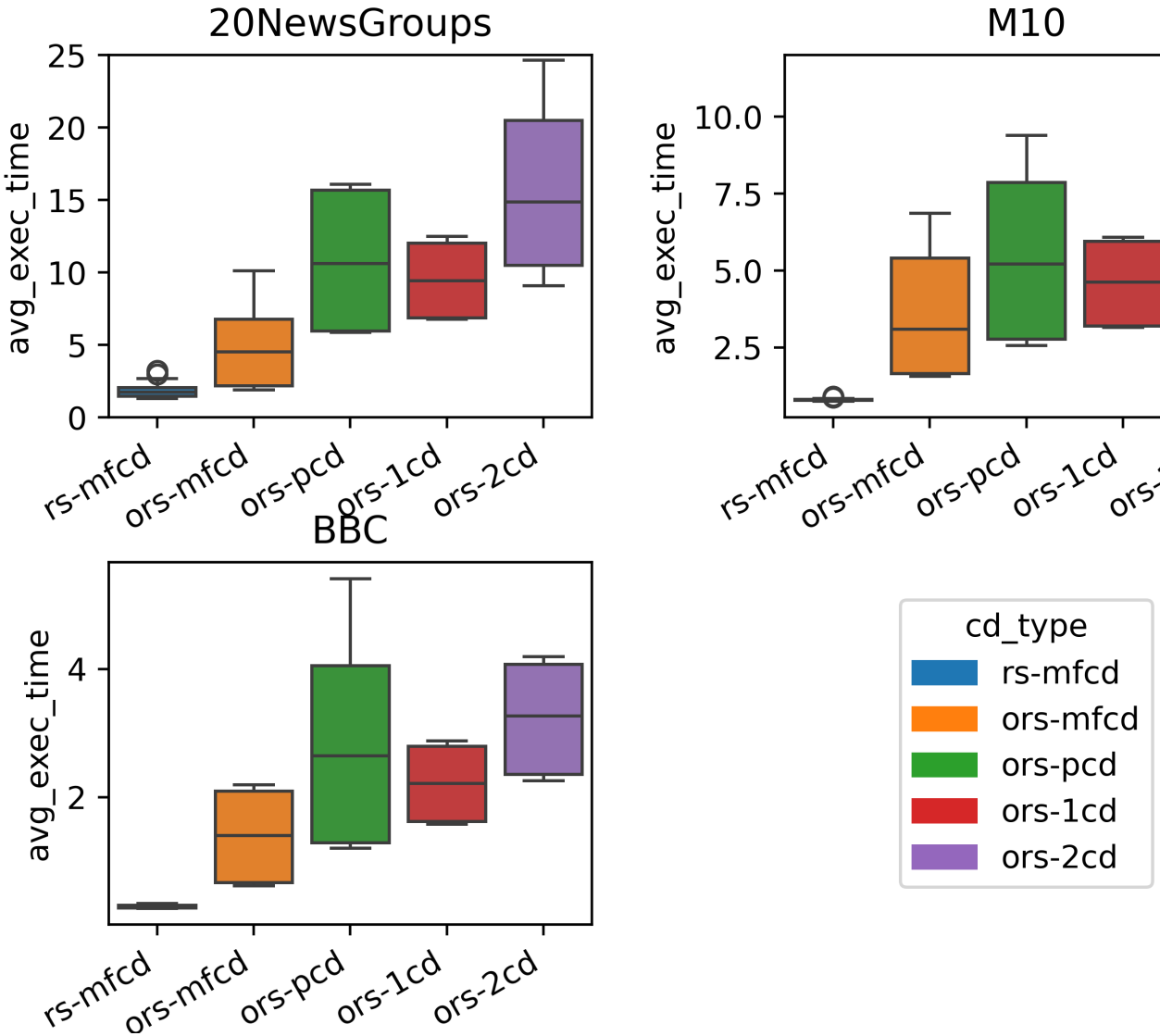
Confronto tra metodi di contrastive divergence

Coherence per Replicated Softmax



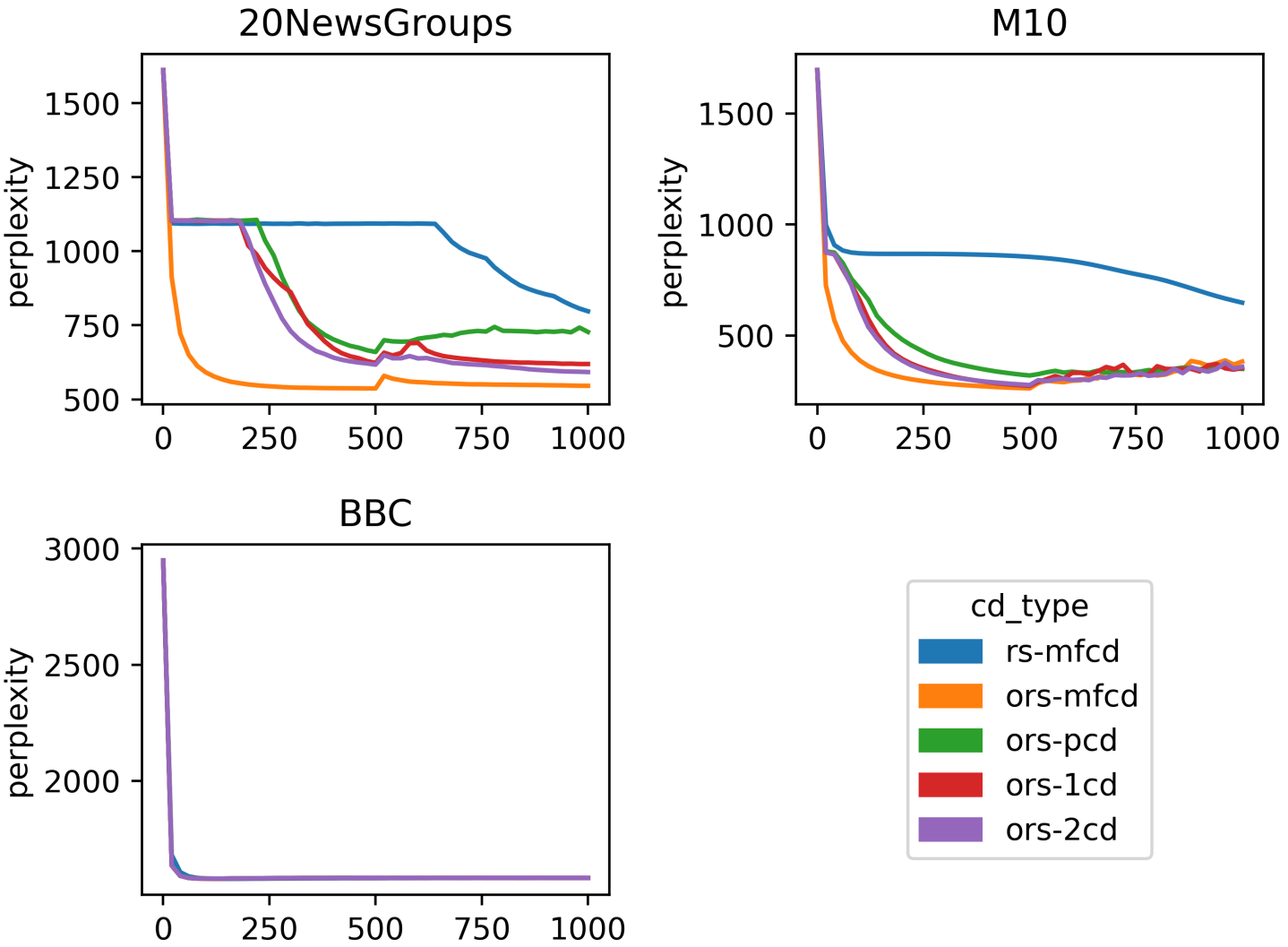
Confronto tra metodi di contrastive divergence

Secondi per iterazione con over Replicated Softmax



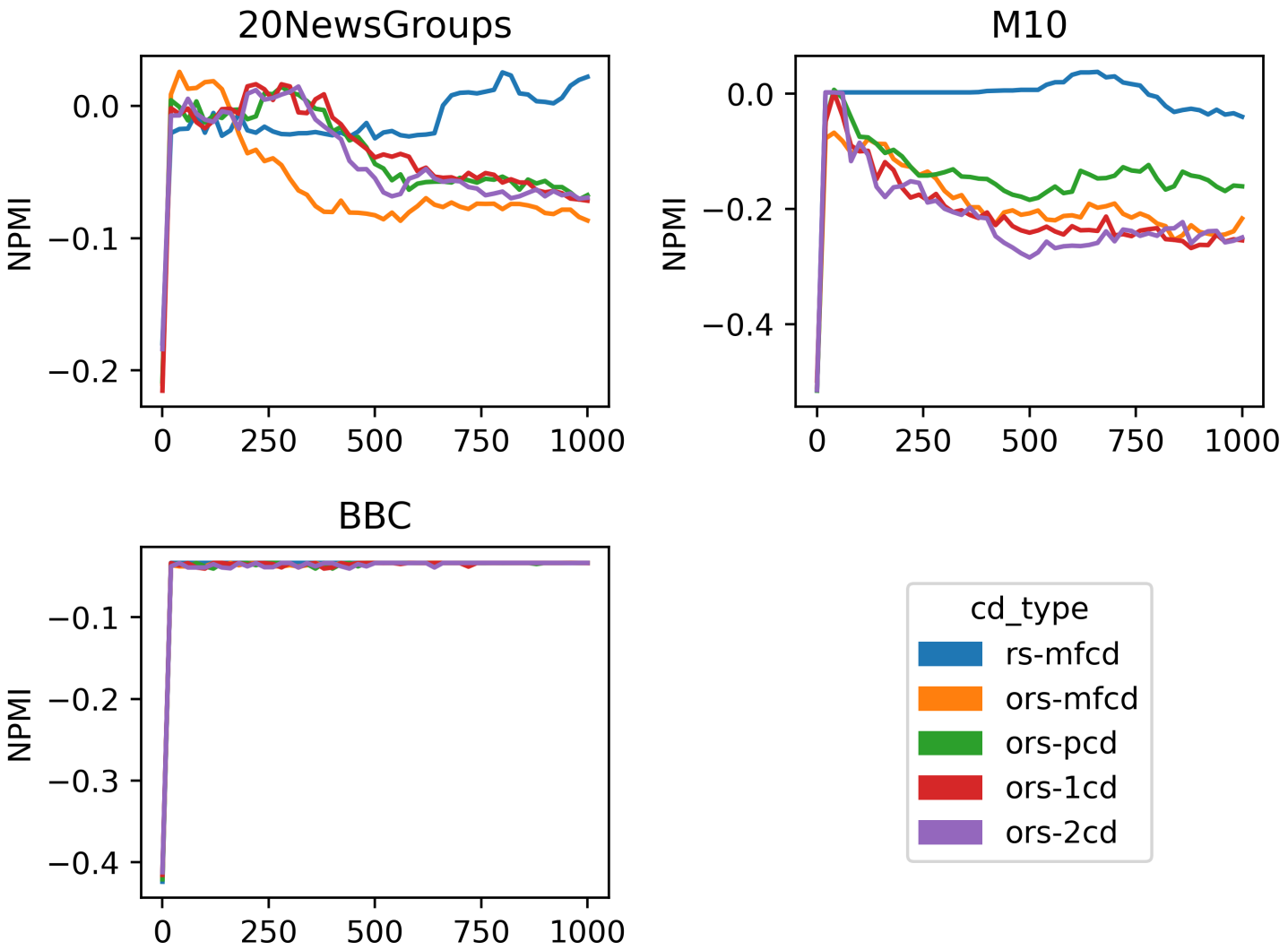
Confronto tra metodi di contrastive divergence

Perplexity per over-Replicated Softmax



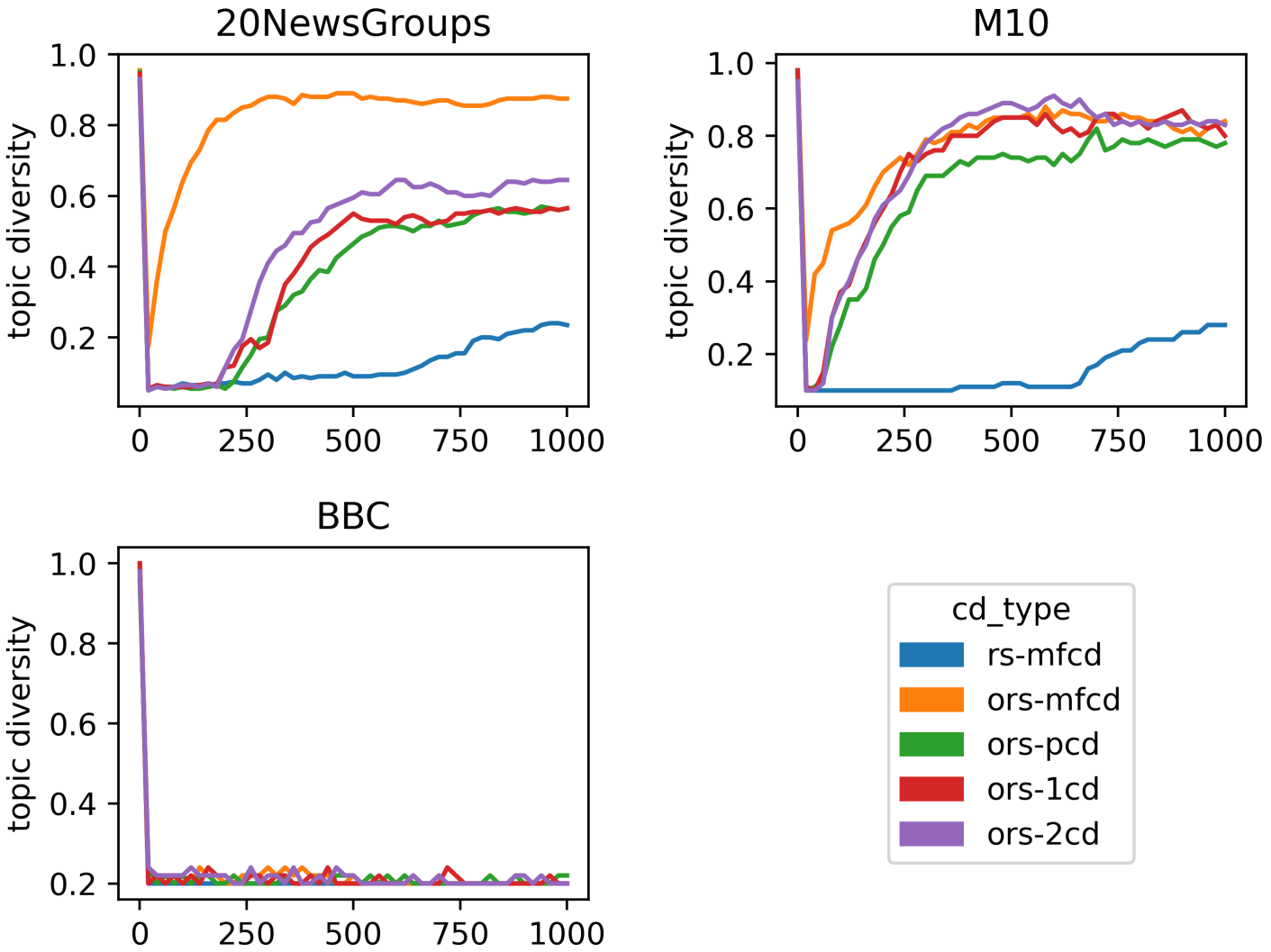
Confronto tra metodi di contrastive divergence

Coherence per over Replicated Softmax



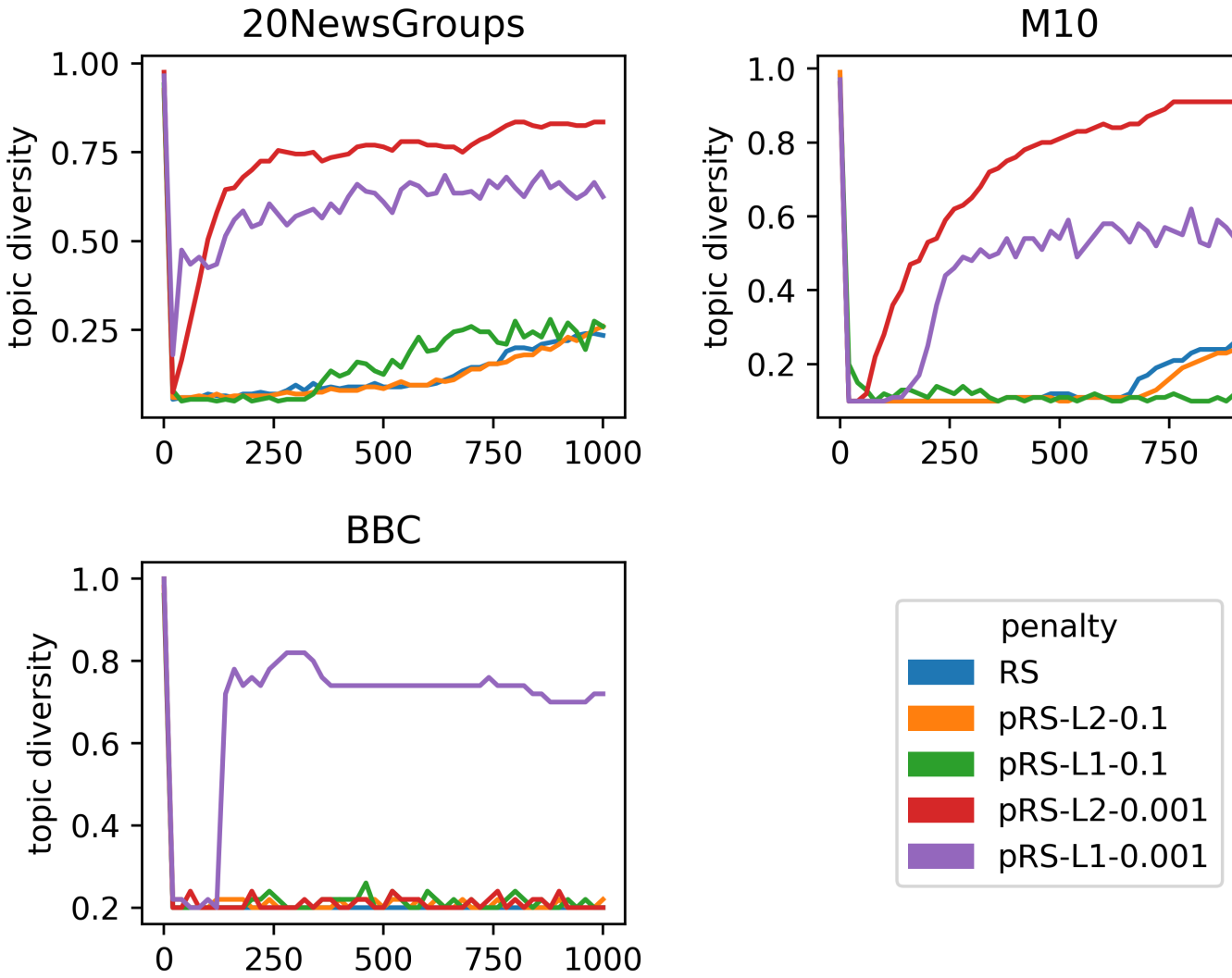
Confronto tra metodi di contrastive divergence

Topic diversity per over Replicated Softmax



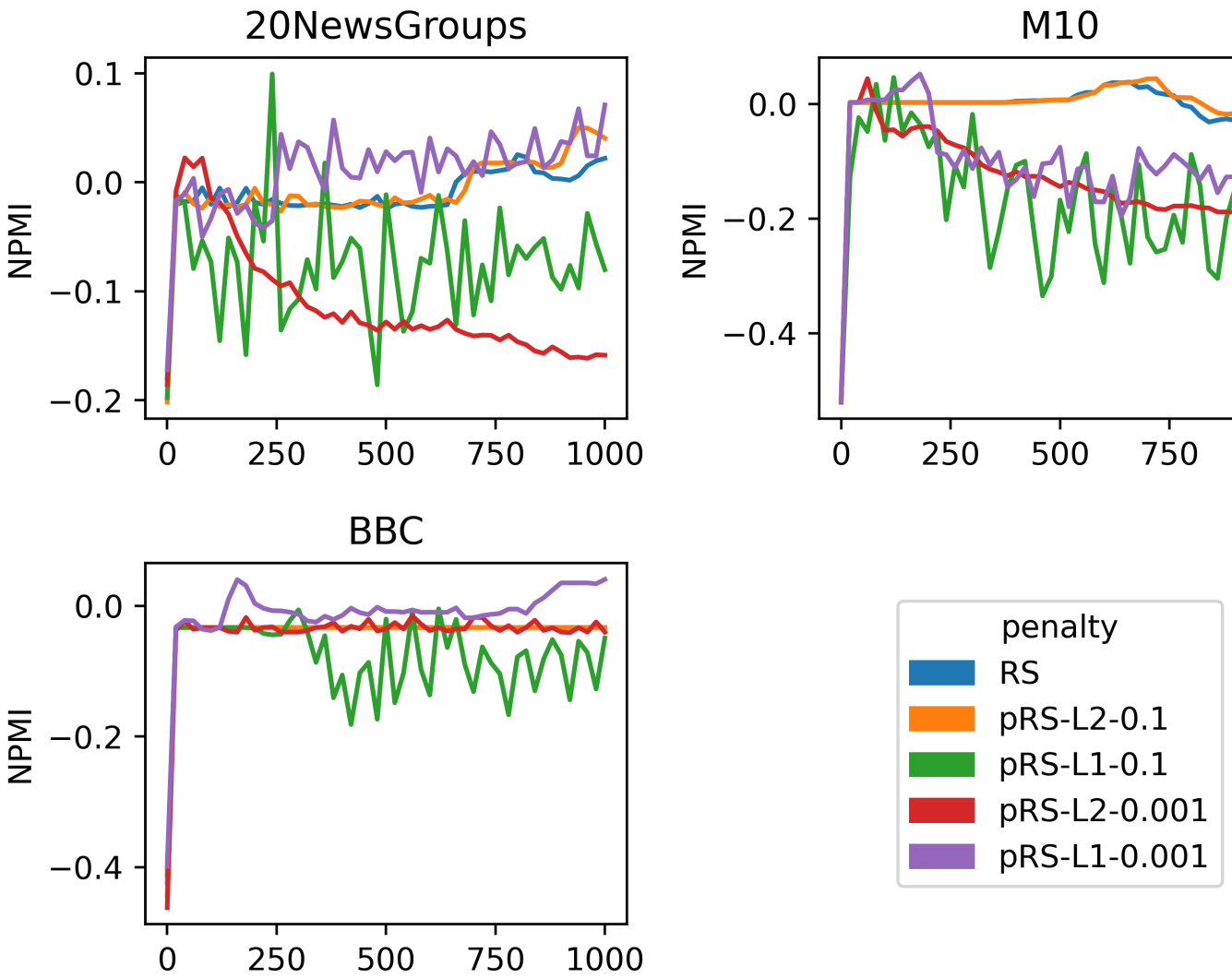
Penalizzazione del modello

Topic Diversity per penalized Replicated Softmax



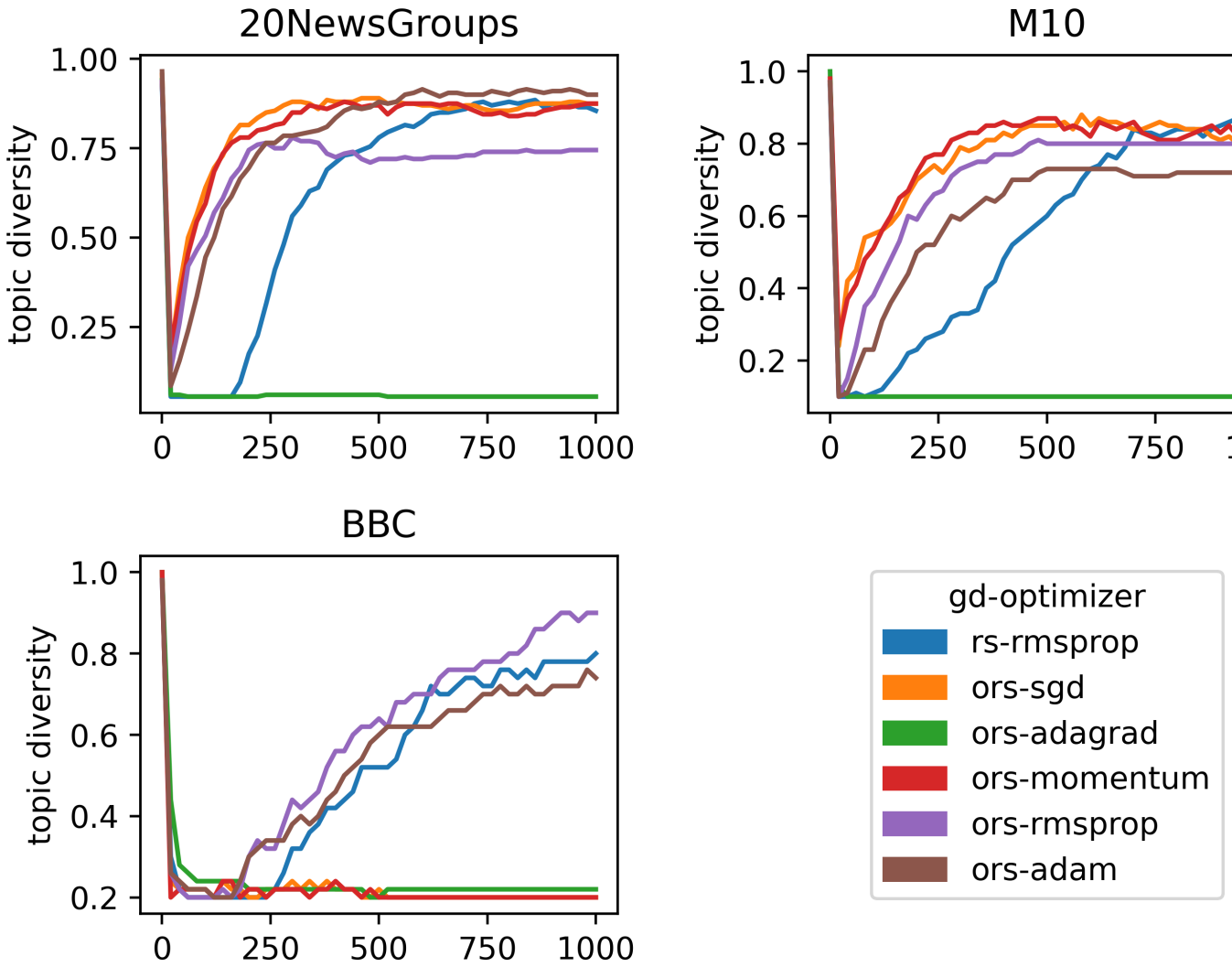
Penalizzazione del modello

Coherence per penalized Replicated Softmax



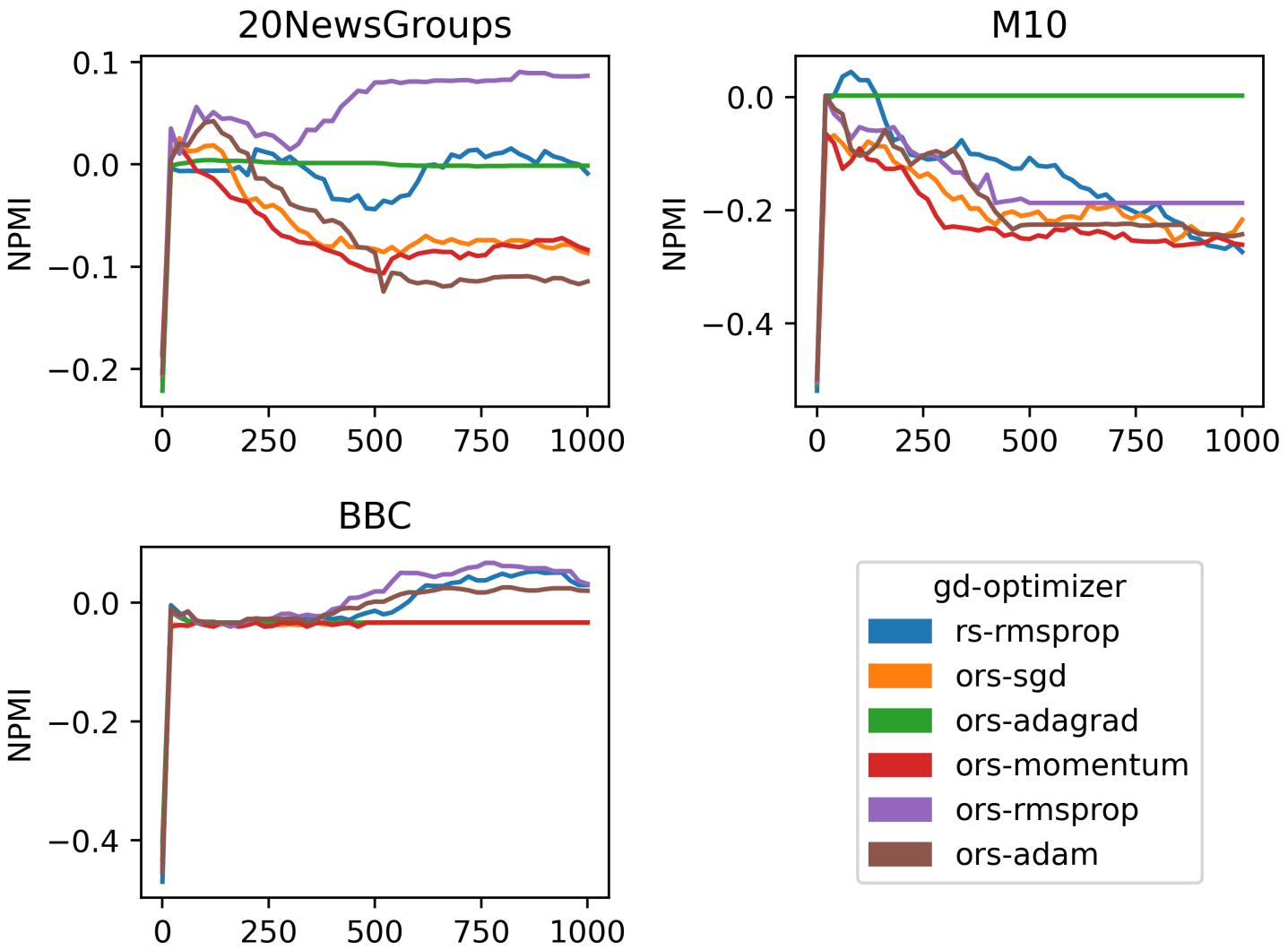
Ottimizzazione del gradiente

Topic Diversity per over Replicated Softmax



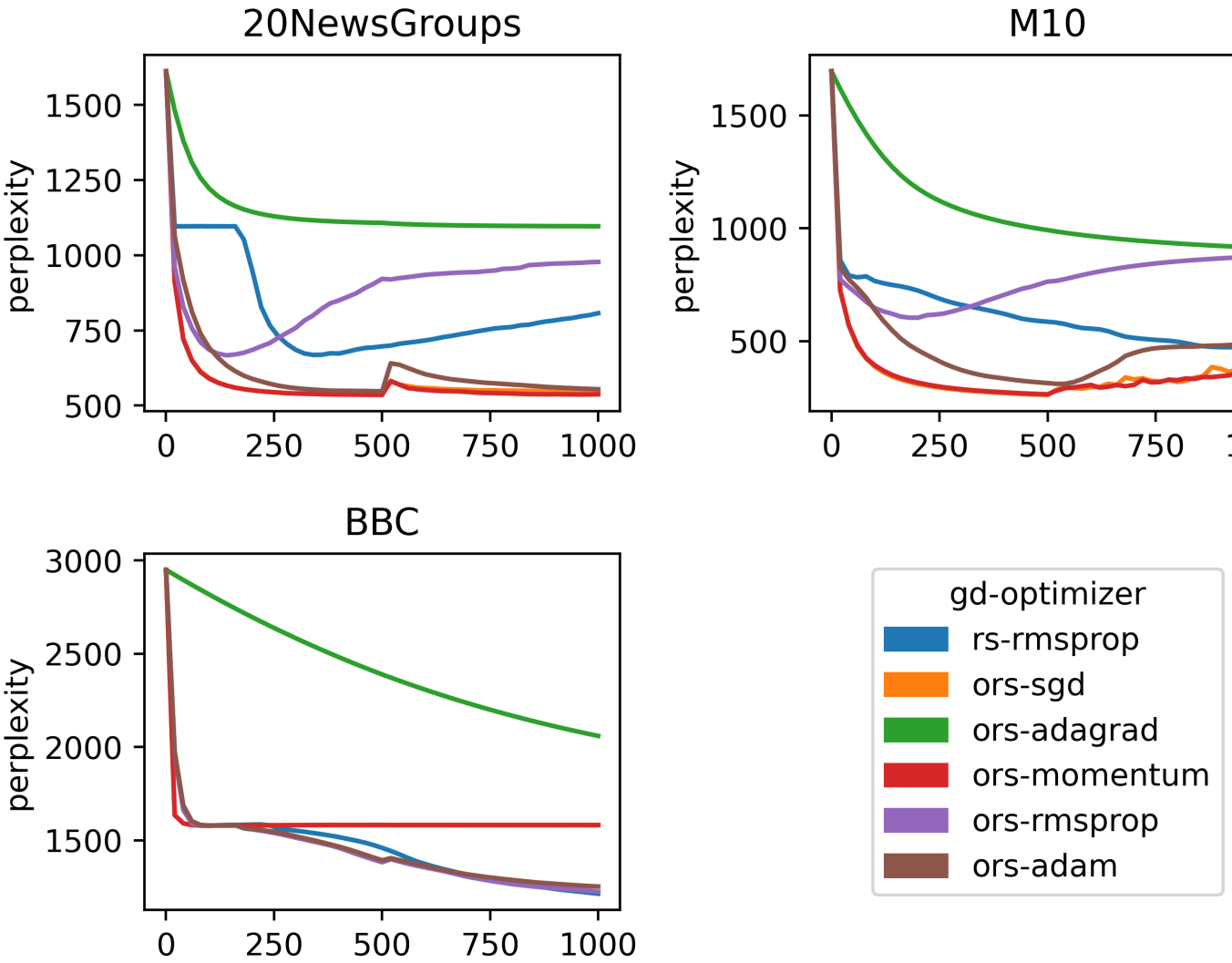
Ottimizzazione del gradiente

Coherence per over Replicated Softmax



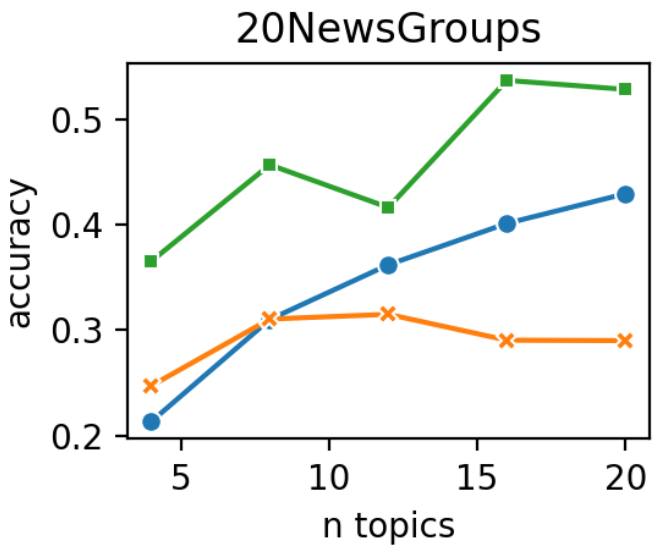
Ottimizzazione del gradiente

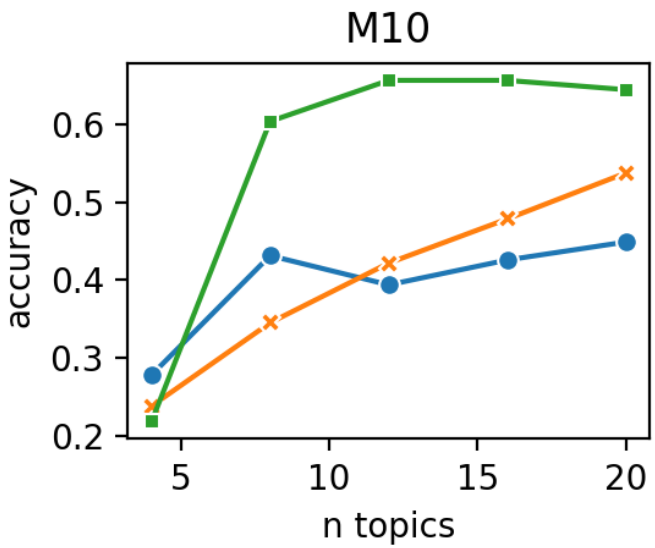
Perplexity per over Replicated Softmax

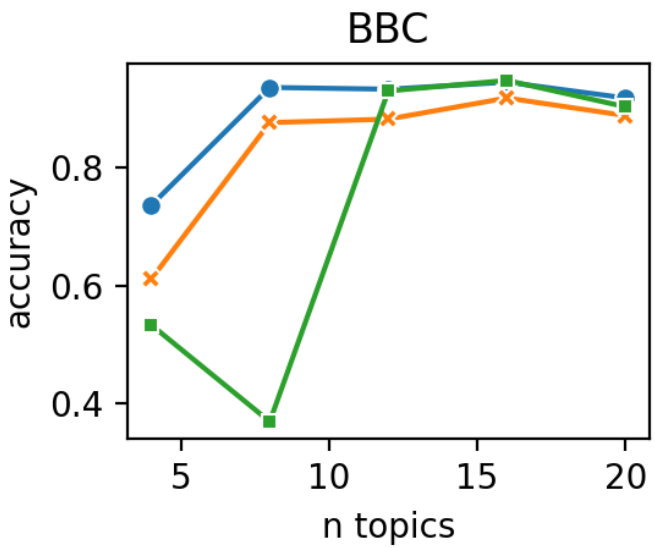


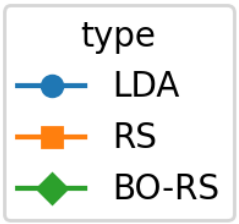
Confronto tra RSM e LDA

Accuracy per numero di topic



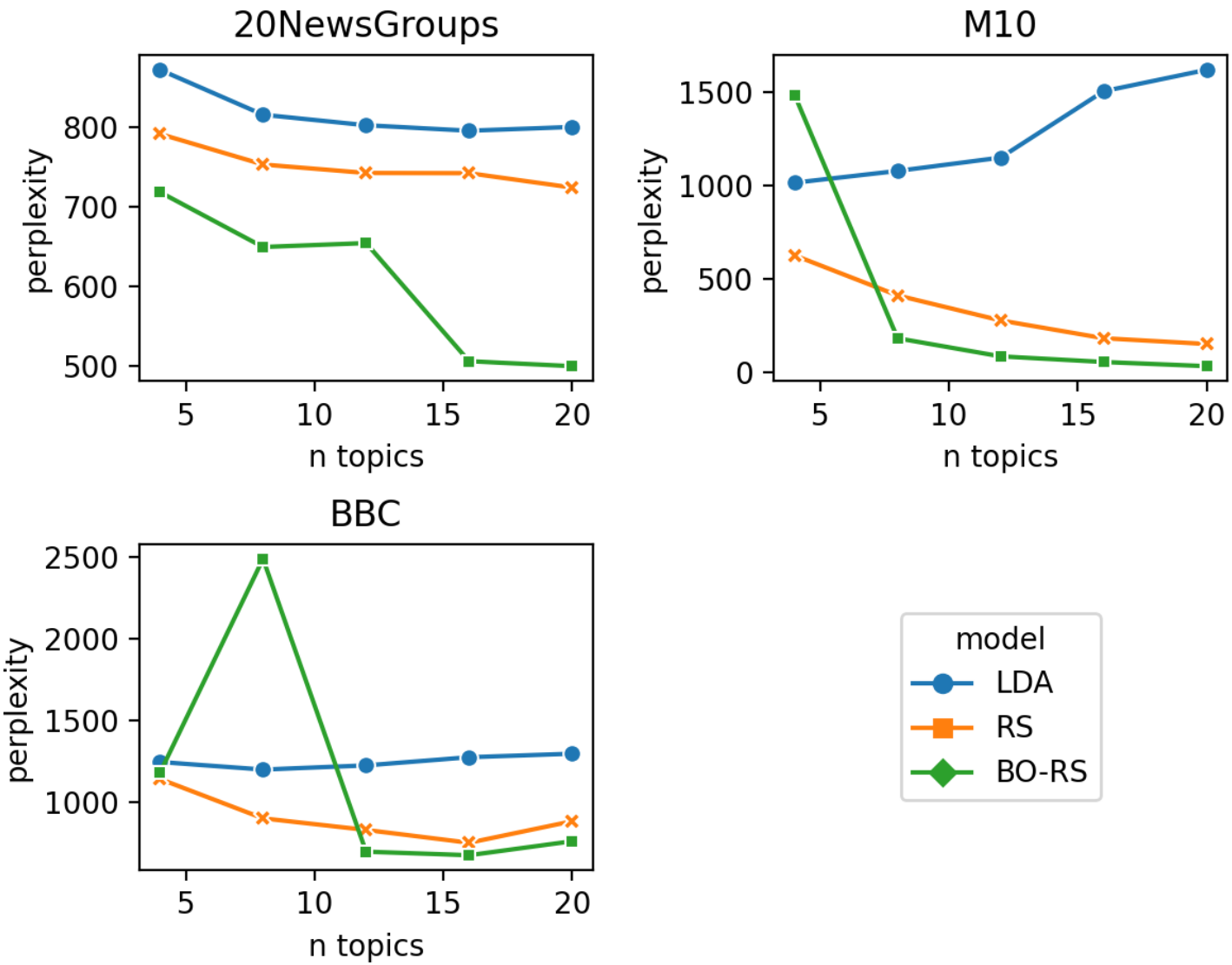






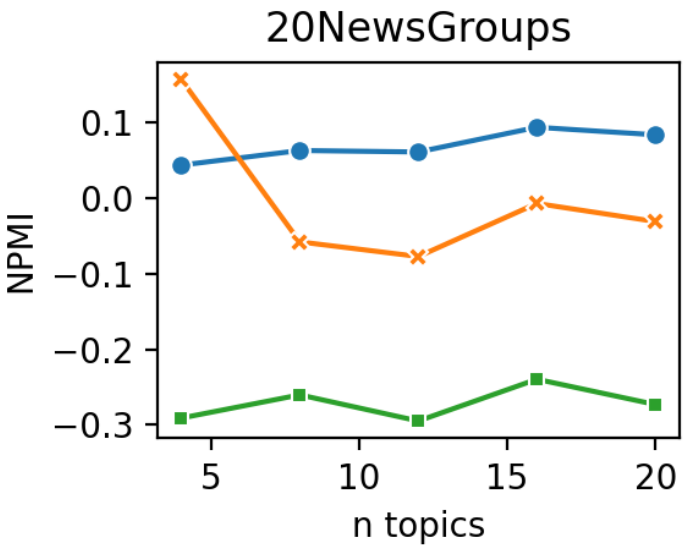
Confronto tra RSM e LDA

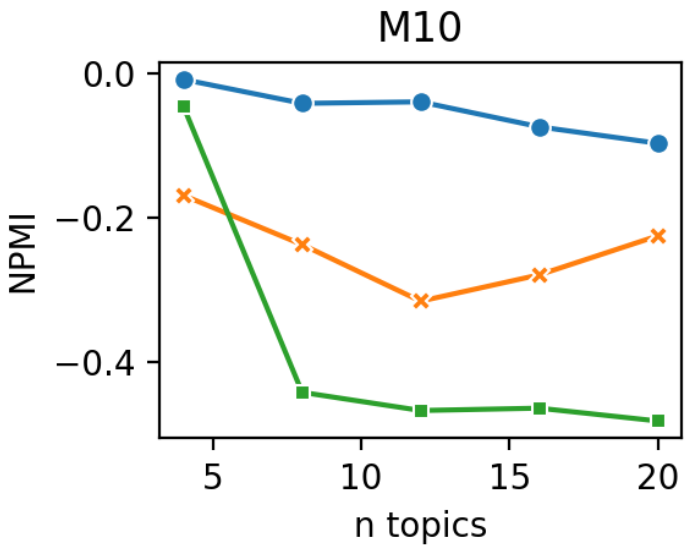
perplexity per numero di topic

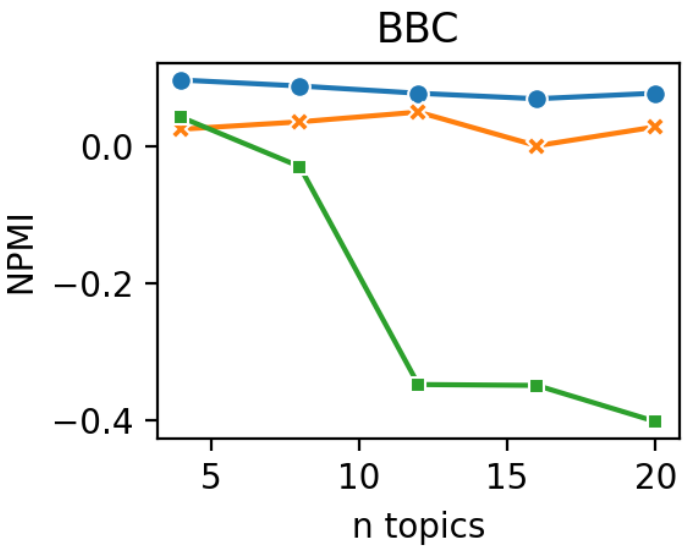


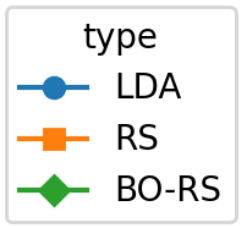
Confronto tra RSM e LDA

Coherence per numero di topic



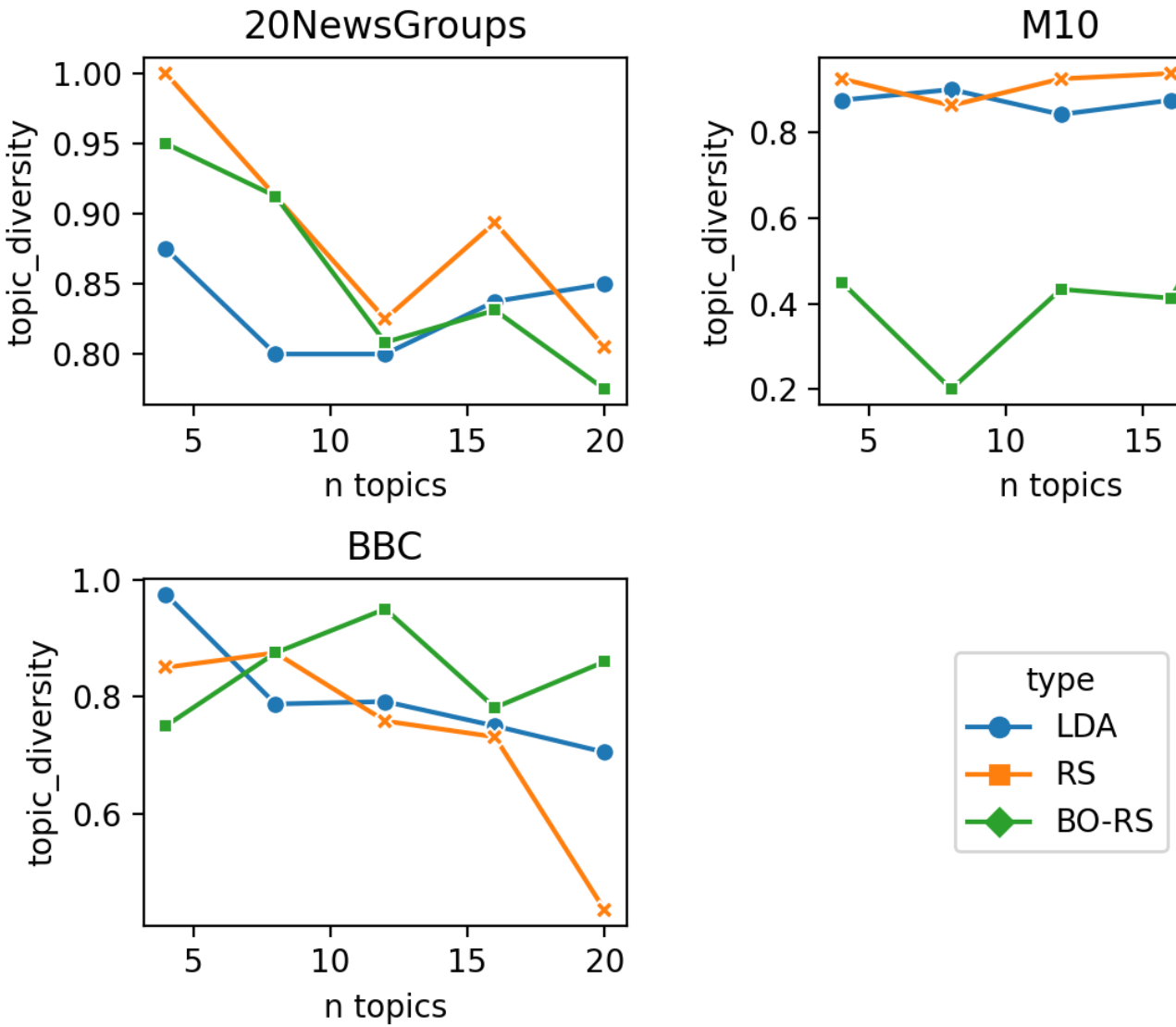






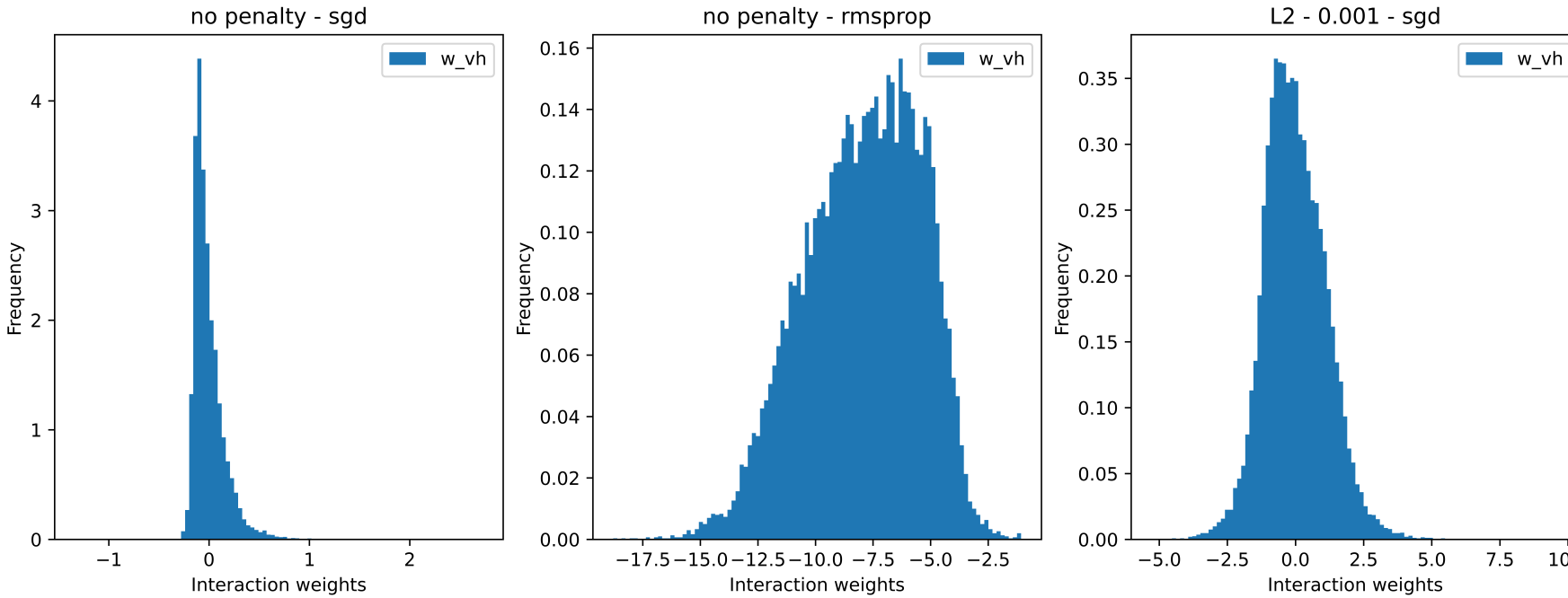
Confronto tra RSM e LDA

Topic diversity per numero di topic



Interpretazione dei pesi di interazione

risultati con Replicated Softmax sul dataset M10



Interpretazione dei pesi di bias

Bias ottenuti con RMSprop: