Non-Convex Projected Gradient Descent for Generalized Low-Rank Tensor Regression

Ganchao Wei

December 1, 2021

Overview

Generalized Tensor Regression

For a scalar response Y, given a covariate tensor $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$, the generalized tensor regression is:

$$p(Y|X, T) = h(Y) \exp\{Y < X, T > -a(< X, T >)\}$$

, where $a(\cdots)$ is a strictly convex and differentiable log-partition function, $h(\cdot)$ is a nuisance parameter, and $T \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is the parameter tensor of interest. Hence The negative log-likelihood risk object is:

$$L(A) = \frac{1}{n} \sum_{i=1}^{n} [a(< X^{(i)}, A >) - Y^{(i)} < X^{(i)}, A > - \log h(Y^{(i)})]$$

Some notations:

- $< \cdot, \cdot >$: inner product
- $||A||_F =< A, A >^{1/2}$
- $||A||_n^2$:empirical norm
- A_A: projection of a tensor A onto A.

Tensor Algebra

There are 2 standard tensor decompositions (e.g. third-order tensors):

- canonical polyadic (CP) decomposition:

$$A = \sum_{k=1}^{r} u_{k,1} \otimes u_{k,2} \otimes u_{k,3}$$

- Tucker decomposition:

$$A_{j_1 j_2 j_3} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} S_{k_1 k_2 k_3} U_{j_1 k_1,1} U_{j_2 k_2,2} U_{j_3 k_3,3}$$

Moreover, there are some tensor operators:

- matricization: $M_m(A)$
- vectorization: $vec(A)$
- slices:

$$\{A_{..j_3} := (A_{j_1 j_2 j_3})_{1 \le j_1 \le d_1, 1 \le j_2 \le d_2} : 1 \le j_3 \le d_3\}$$

Low-dimensional Structural Assumption

Consider 3 (non-convex) specific examples of low-rank structure:

- Low-rank structure on the matrix slices:

$$\Theta_1(r) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \sum_{j_3=1}^{d_3} \operatorname{rank}(A_{\cdot\cdot j_3}) \leq r \right\}$$

- Bound the maximum of the rank of each slice and sparsity along the matrix slices:

$$\Theta_2(r,s) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \max_{j_3} \operatorname{rank}(A_{\cdot\cdot j_3}) \leq r, \sum_{j_3=1}^{d_3} \mathbb{I}(A_{\cdot\cdot j_3} \neq 0) \leq s \right\}$$

- Tucker ranks are upper bounded:

$$\Theta_3(r_1,r_2,r_3) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : r_i(A) \leq r_i \text{ for } i = 1,2,3 \right\}$$

Projected Gradient Descent

Minimize a general loss function $f(A)$ subject to constraint $A \in \Theta$:

Algorithm 1 Projected Gradient Descent

1: **Input :** data $\mathbf{Y}, \mathbf{X}$, parameter space Θ, iterations K, step size η
2: **Initialize :** $k = 0$, $\widehat{T}_0 \in \Theta$
3: **for** $k = 1, 2, \ldots, K$ **do**
4: $\quad g_k = \widehat{T}_k - \eta \nabla f(\widehat{T}_k)$ (gradient step)
5: $\quad \widehat{T}_{k+1} = P_\Theta(g_k)$ or $\widehat{T}_{k+1} = \widehat{P}_\Theta(g_k)$ ((approximate) projection step)
6: **end for**
7: **Output :** $\widehat{T}_K$

Denote 2 specific projections as:

- projection to the set of s-sparse vectors:

$$\tilde{P}_s(v) := \underset{\|z\|_{\ell_0} \le s}{\arg\min} \|z - v\|_{\ell_2}$$

- rank-r projection of a matrix:

$$\bar{P}_r(M) := \underset{\mathrm{rank}(Z) \le r}{\arg\min} \|Z - M\|_{\mathrm{F}}$$

Properties for Θ and its projection

View Θ as a member of a collection of subspaces $\{\Theta(t) : t \in \Xi\}$. Then the 3 low-dimensional structures:

- $\Theta_1(r)$:

$$\Xi = \{0, \ldots, d_3 \cdot \min\{d_1, d_2\}\}$$

- $\Theta_2(r, s)$:

$$\Xi = \{(0, 0), \ldots, (\min\{d_1, d_2\}, d_3)\}$$

- $\Theta_3(r_1, r_2, r_3)$:

$$\Xi = \{(0, 0, 0), \ldots, (\min\{d_1, d_2 d_3\}, \min\{d_2, d_1 d_3\}, \min\{d_3, d_1 d_2\})\}$$

Properties for Θ and its projection

Further, there are some definitions:

Definition 1. A set $\{\Theta(t) : t \in \Xi\}$ is a *superadditive and partially ordered collection of symmetric cones* if

(1) each member $\Theta(t)$ is a *symmetric cone* in that if $z \in \Theta(t)$, then $cz \in \Theta(t)$ for any $c \in \mathbb{R}$;

(2) the set is *partially ordered* in that for any $t_1 \leq t_2$, $\Theta(t_1) \subset \Theta(t_2)$;

(3) the set is *superadditive* in that $\Theta(t_1) + \Theta(t_2) \subset \Theta(t_1 + t_2)$.

And the definition of contractive projection:

Definition 2. We say that a set $\{\Theta(t) : t \geq 0\}$ and corresponding operators $Q_{\Theta(t)} : \cup_t \Theta(t) \mapsto \Theta(t)$ satisfy the *contractive projection property* for some $\delta > 0$, denoted by $\mathrm{CPP}(\delta)$, if for any $t_1 < t_2 < t_0$, $Y \in \Theta(t_1)$, and $Z \in \Theta(t_0)$:

$$\|Q_{\Theta(t_2)}(Z) - Z\|_{\mathrm{F}} \leq \delta \left\| \frac{t_0 - t_2}{t_0 - t_1} \right\|_{\ell_\infty}^{1/2} \cdot \|Y - Z\|_{\mathrm{F}}.$$

Here, when $\Theta(t)$ is indexed by multi-dimensional t, the division $\frac{t_0 - t_2}{t_0 - t_1}$ refers to a vector with the jth element to be $\frac{(t_0)_j - (t_2)_j}{(t_0)_j - (t_1)_j}$.

Restricted strong convexity

To guarantee the PGD performance, we further need the restricted strong convexity and smoothness conditions (RSCS):

Definition 3. We say that a function f satisfies *restricted strong convexity and smoothness conditions* $RSCS(\Theta, C_l, C_U)$ for a set Θ, and $0 < C_l < C_u < \infty$ if for any $A \in \Theta$, $\nabla^2 f(A)$ is positive semidefinite such that for any $B \in \Theta$

$$C_l \cdot \|B\|_{\mathrm{F}} \leq \|\nabla^2 f(A) \cdot \mathrm{vec}(B)\|_{\ell_2} \leq C_u \cdot \|B\|_{\mathrm{F}},$$

for some constants $C_l < C_u$. Note that $RSCS(\Theta, C_l, C_U)$ reduces to restricted strong convexity and restricted smoothness assumptios for vectors and matrices (see e.g. Jain et al. (2014, 2016)) when A and B are vectors and matrices. We first state the following Theorem about the PGD performance under general loss function which is a tensor version of the results in Jain et al. (2014, 2016).

Restricted strong convexity

Under this condition, we can give the error bound for general loss function:

Theorem 1 ***(PGD Error Bound for General Loss Function)*** *Suppose that* $\{\Theta(t) : t \geq 0\}$ *is a superadditive and partially ordered collection of symmetric cones, together with operators* $\{P_{\Theta(t)} : t \geq 0\}$ *which obey CPP*(δ) *for some constant* $\delta > 0$*, and* f *satisfies RSCS*$(\Theta(t_0), C_l, C_u)$ *for some constants* C_l *and* C_u*. Let* $\widehat{T}_K$ *be the output from the* K*th iteration of applying PGD algorithm with step size* $\eta = 1/C_u$*, and projection* $P_{\Theta(t_1)}$ *where*

$$t_1 = \left\lceil \frac{4\delta^2 C_u^2 C_l^{-2}}{1 + 4\delta^2 C_u^2 C_l^{-2}} \cdot t_0 \right\rceil .$$

Then

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq 4\eta C_u C_l^{-1} \cdot \sup_{A \in \Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)} \langle \nabla f(T), A \rangle + \epsilon,$$

for any

$$K \geq 2 C_u C_l^{-1} \log \left(\frac{\|T\|_{\mathrm{F}}}{\epsilon} \right).$$

We then specialize f to the generalized linear model.

Generalized linear models

For convenience, we assume & define:

- Gaussian design og independent samples tensors:

$$\mathrm{vec}(X^{(i)}) \sim \mathcal{N}(0, \Sigma) \text{ where } \Sigma \in \mathbb{R}^{D_N \times D_N}. \tag{4}$$

- Σ has bounded eigenvalues:

$$c_\ell^2 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_u^2, \tag{5}$$

- Gaussian width:

A quantity that emerges from our analysis is the *Gaussian width* (see, e.g., Gordon, 1988) of a set $S \subset \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ which is defined to be:

$$w_G(S) := \mathbb{E}\left(\sup_{A \in S} \langle A, G \rangle\right),$$

Generalized linear models

Lemma 1 *Assume that (4) and (5) hold. For any $\tau > 1$, there exist constants $c_1, c_2, c_3 > 0$ such that if $n \geq c_1 w_G^2[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)]$, then with probability at least $1 - c_2 \exp(-c_3 w_G^2[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)])$,*

$$\left(\tau^{-1} c_l\right)^2 \|A\|_{\mathrm{F}}^2 \leq \frac{1}{n} \sum_{i=1}^{n} \langle X^{(i)}, A \rangle^2 \leq (\tau c_u)^2 \|A\|_{\mathrm{F}}^2, \qquad \forall A \in \Theta.$$

Theorem 2 ***(PGD Error Bound for Generalized Linear Model)*** *Suppose that $\{\Theta(t) : t \geq 0\}$ is a superadditive and partially ordered collection of symmetric cones, and together with operators $\{P_{\Theta(t)} : t \geq 0\}$ which obey CPP(δ) for some constant $\delta > 0$. Assume that $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$ follow the generalized linear model (1) and $X^{(i)}$'s satisfy (4) and (5), $\mathbb{E}|Y^{(i)}|^q \leq M_Y$ for some $q > 2$ and $M_Y > 0$, $1/\tau_0^2 \leq Var(Y^{(i)}) \leq \tau_0^2$ for $i = 1, \ldots, n$ and some $\tau_0 > 0$, and $n > c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$ for some t_0 and $c_1 > 0$. Let $\widehat{T}_K$ be the output from the K^{th} iteration of applying PGD algorithm to (2) with step size $\eta = (\tau c_u)^{-2}$ and projection $P_{\Theta(t_1)}$ where*

$$t_1 = \left\lceil \frac{4\delta^2 \tau^8 \kappa^4}{1 + 4\delta^2 \tau^8 \kappa^4} \cdot t_0 \right\rceil,$$

for any given $\tau > \tau_0$. Then there exist constants $c_2, c_3, c_4, c_5 > 0$ such that

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq \frac{c_5 \eta \tau^4 \kappa^2 c_u M_Y^{1/q}}{\sqrt{n}} \cdot w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] + \epsilon,$$

with probability at least

$$1 - K c_2 \exp\left\{-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]\right\} - K c_4 n^{-(q/2-1)} \log^q n,$$

for any

$$K \geq 2\tau^4 \kappa^2 \log\left(\frac{\|T\|_{\mathrm{F}}}{\epsilon}\right).$$

Normal linear regression

The moment conditions ensure the restricted strong convexity & restricted smoothness conditions. When considering normal linear regression, these conditions could be further removed.

Theorem 3 ***(PGD Error Bound for Normal Linear Regression)*** *Suppose that* $\{\Theta(t) : t \geq 0\}$ *is a superadditive and partially ordered collection of symmetric cones, and together with operators* $\{P_{\Theta(t)} : t \geq 0\}$ *which obey CPP*(δ) *for some constant* $\delta > 0$. *Assume that* $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$ *follow the Gaussian linear model* (6) *where* $n > c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$ *for some* t_0 *and* $c_1 > 0$. *Let* $\widehat{T}_K$ *be the output from the* K^{th} *iteration of applying PGD algorithm to* (2) *with step size* $\eta = (\tau c_u)^{-2}$ *and projection* $P_{\Theta(t_1)}$ *where*

$$t_1 = \left\lceil \frac{4\delta^2\tau^8\kappa^4}{1 + 4\delta^2\tau^8\kappa^4} \cdot t_0 \right\rceil,$$

for any given $\tau > 1$. *Then there exist constants* $c_2, c_3 > 0$ *such that*

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq \frac{8\eta\tau^4\kappa^2 c_u \sigma}{\sqrt{n}} w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] + \epsilon,$$

with probability at least

$$1 - K c_2 \exp\left\{-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]\right\},$$

for any

$$K \geq 2\tau^4\kappa^2 \log\left(\frac{\|T\|_{\mathrm{F}}}{\epsilon}\right).$$

Normal linear regression

The convex regularization approach:

$$\widehat{T} \in \underset{A \in \mathbb{R}^{d_1 \times \cdots \times d_N}}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)} \rangle\|_{\mathrm{F}}^2 + \lambda \mathcal{R}(A) \right\}, \tag{8}$$

The error bound:

$$\max\left\{ \|\widehat{T} - T\|_n, \|\widehat{T} - T\|_{\mathrm{F}} \right\} \lesssim \frac{\sqrt{s(\Theta)}\lambda}{\sqrt{n}}$$

when $\lambda = 2\omega_G(\mathbb{B}_R(1))$

$$\max\left\{ \|\widehat{T} - T\|_n, \|\widehat{T} - T\|_{\mathrm{F}} \right\} \lesssim \frac{\sqrt{s(\Theta)}w_G(\mathbb{B}_{\mathcal{R}}(1))}{\sqrt{n}}$$

Normal linear regression

By

$$
\begin{aligned}
w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] &= \mathbb{E}\Big[\sup_{A \in \Theta(t_0), \|A\|_{\mathrm{F}} \leq 1} \langle A, G \rangle\Big] \\
&\leq \mathbb{E}\Big[\sup_{\mathcal{R}(A) \leq \sqrt{s(\Theta(t_0))}} \langle A, G \rangle\Big] \\
&= \sqrt{s(\Theta(t_0))}\mathbb{E}\Big[\sup_{\mathcal{R}(A) \leq 1} \langle A, G \rangle\Big] = \sqrt{s(\Theta(t_0))} w_G[\mathbb{B}_{\mathcal{R}}(1)].
\end{aligned}
$$

The non-convex error bound is always no larger than the convex error bound.

Specific low rank structure

$\Theta_1(r)$ is identical to the case of low rank matrix estimation. For $\Theta_2(r, s)$:

Lemma 2 *Let the projection operator $P_{\Theta_2(r,s)}$ be defined above. Suppose $Z \in \Theta_2(r_0, s_0)$, and $r_1 < r_2 < r_0, s_1 < s_2 < s_0$. Then for any $Y \in \Theta_2(r_1, s_1)$, we have*

$$\|P_{\Theta_2(r_2,s_2)}(Z) - Z\|_{\mathrm{F}} \leq (\alpha + \beta + \alpha\beta) \cdot \|Y - Z\|_{\mathrm{F}}.$$

where $\alpha = \sqrt{(s_0 - s_2)/(s_0 - s_1)}$, $\beta = \sqrt{(r_0 - r_2)/(r_0 - r_1)}$.

Consequently we have the following Theorem:

Theorem 4 *Let $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ follow a Gaussian linear model as defined by (6) with $T \in \Theta_2(r, s)$ and*

$$n \geq c_1 \cdot sr(d_1 + d_2 + \log d_3)$$

for some constant $c_1 > 0$. Then, applying the PGD algorithm with step size $\eta = (\tau c_u)^{-2}$ and projection $P_{\Theta(r',s')}$ where

$$s' = \lceil 36\tau^8\kappa^4 s \rceil, \qquad \text{and} \qquad r' = \lceil 36\tau^8\kappa^4 r \rceil,$$

guarantees that, with probability at least $1 - Kc_2 \exp\{-c_3 \max(d_1, d_2, \log d_3)\}$, after $K \geq 2\tau^4\kappa^2 \log(\|T\|_{\mathrm{F}}/\epsilon)$ iterations,

$$\|\widehat{T}_K - T\|_{\mathrm{F}} \leq c_4\sigma\sqrt{\frac{sr \max\{d_1, d_2, \log(d_3)\}}{n}} + \epsilon$$

for any $\tau > 1$, and some constants $c_2, c_3, c_4 > 0$.

Specific low rank structure

The $\Theta_2(r, s)$ is not suitable for convex regularization, so only compare for $\Theta_1(r)$.

$$\|\widehat{T} - T\|_{\mathrm{F}} \lesssim \sqrt{\frac{r \max(d_1, d_2, \log d_3)}{n}}.$$

Notice that both $\Theta_1(r)$ and $\Theta_2(r, s)$ focus on the low-rankness of matrix slices of a tensor, and actually $\Theta_1(\cdot)$ can be seen as relaxation of $\Theta_2(\cdot, \cdot)$ since $\Theta_2(s, r) \subset \Theta_1(sr)$. Theorem 4 guarantees that, under the restriction of sparse slices of low-rank matrices, PGD achieves the linear convergence rate with the statistical error of order

$$\sqrt{\frac{sr \max\{d_1, d_2, \log(d_3)\}}{n}}.$$

If we compare this result with the risk bound of the convex regularization approach where the true tensor parameter lies in $\Theta_1(r)$ we see that replacing r by sr yields the same rate which makes intuitive sense in light of the observation that $\Theta_2(s, r) \subset \Theta_1(sr)$.

Specific low rank structure

We further consider the low Tucker rank $\Theta_3(r_1, r_2, r_3)$:

Lemma 3 *Suppose* $Z \in \Theta_3(r_1^{(0)}, r_2^{(0)}, r_3^{(0)})$, *and* $r_i^{(1)} < r_i^{(2)} < r_i^{(0)}$ *for* $i = 1, 2, 3$. *Then for any* $Y \in \Theta_3(r_1^{(1)}, r_2^{(1)}, r_3^{(1)})$, *we have*

$$\|\widehat{P}_{\Theta_3(r_1,r_2,r_3)}(Z) - Z\|_{\mathrm{F}} \leq [(\beta_1+1)(\beta_2+1)(\beta_3+1) - 1]\|Y - Z\|_{\mathrm{F}}$$

where $\beta_i = \sqrt{(r_i^{(0)} - r_i^{(2)})/(r_i^{(0)} - r_i^{(1)})}$.

This allows us to derive the following result for the PGD algorithm applied with projection operator $\widehat{P}_{\Theta_3(r_1',r_2',r_3')}(\cdot)$. Basically, the sequential matrix low-rank projection, as an approximate projection onto low Tucker rank subset, could achieve the same order error rate as the exact low Tucker rank projection which might involve expensive iterative computation.

Theorem 5 *Let* $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ *follow a Gaussian linear model as defined by* (6) *with* $T \in \Theta_3(r_1, r_2, r_3)$ *and*

$$n \geq c_1 \cdot \min\{r_1(d_1 + d_2d_3), r_2(d_2 + d_1d_3), r_3(d_3 + d_1d_2)\},$$

for some constant $c_1 > 0$. *Then, applying the PGD algorithm with step size* $\eta = (\tau c_u)^{-2}$ *and projection* $\widehat{P}_{\Theta_3(r_1',r_2',r_3')}$ *where*

$$r_i' = \lceil 196\tau^8\kappa^4 r_i \rceil \text{ for } i = 1, 2, 3$$

guarantees that, with probability at least $1 - Kc_2 \exp\{-c_3 \min(d_1+d_2d_3, d_2+d_1d_3, d_3+d_1d_2)\}$, *after* $K \geq 2\tau^4\kappa^2 \log(\|T\|_{\mathrm{F}}/\epsilon)$ *iterations,*

$$\|\widehat{T}_K - T\|_{\mathrm{F}} \leq c_4\sigma\sqrt{\frac{\min\{r_1(d_1 + d_2d_3), r_2(d_2 + d_1d_3), r_3(d_3 + d_1d_2)\}}{n}} + \epsilon$$

for any $\tau > 1$, *and some constants* $c_2, c_3, c_4 > 0$.

Specific low rank structure

The error bound for convex regularization:

$$\|\widehat{T} - T\|_{\mathrm{F}} \lesssim \sqrt{\frac{\max(r_1, r_2, r_3) \cdot \max(d_1 + d_2 d_3, d_2 + d_1 d_3, d_3 + d_1 d_2)}{n}}.$$

Notice that min is replaced by max, so PGD is better.

Simulations

How to select step-size for PGD?

- if η is too large: divergence
- if η is too small: slow convergence

2 directions:

- start with large, decrease when divergence is observed
- start with small, increase but step back when divergence is observed.

For all simulations, generally we can see:

- larger $\eta \Rightarrow$ faster convergence
- larger r' or $s' \Rightarrow$ faster convergence: wrong rank/sparsity will do harm to the performance of PGD.

Simulations

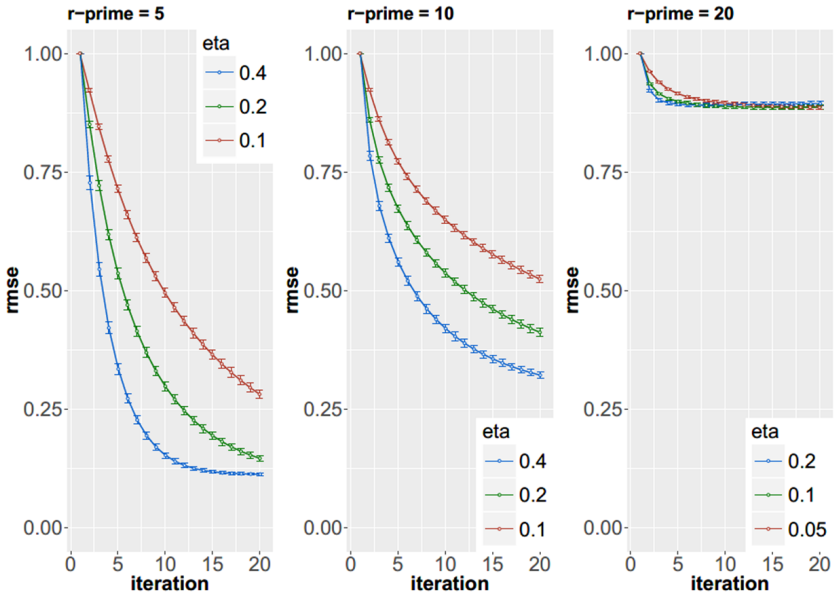

Figure 1: Case 1a: Low CP rank

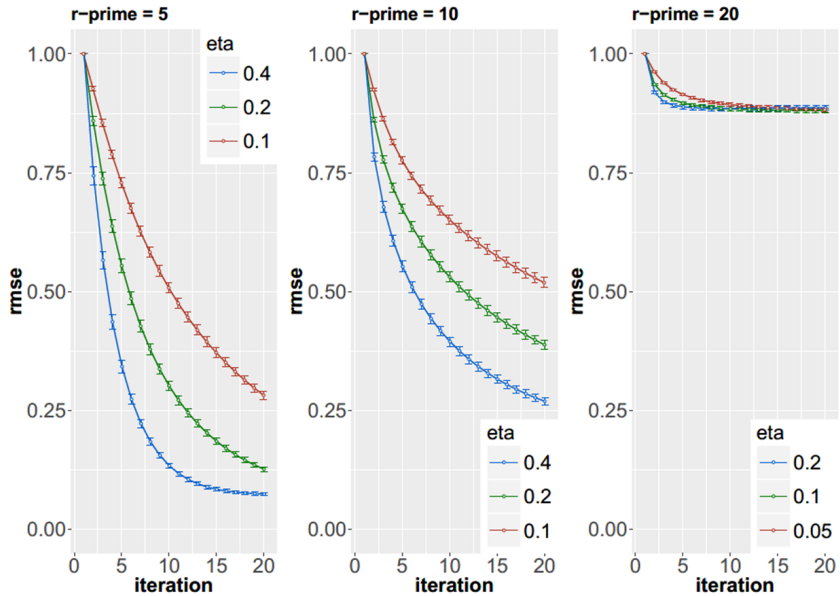

Figure 2: Case 2a: Low Tucker rank

Simulations

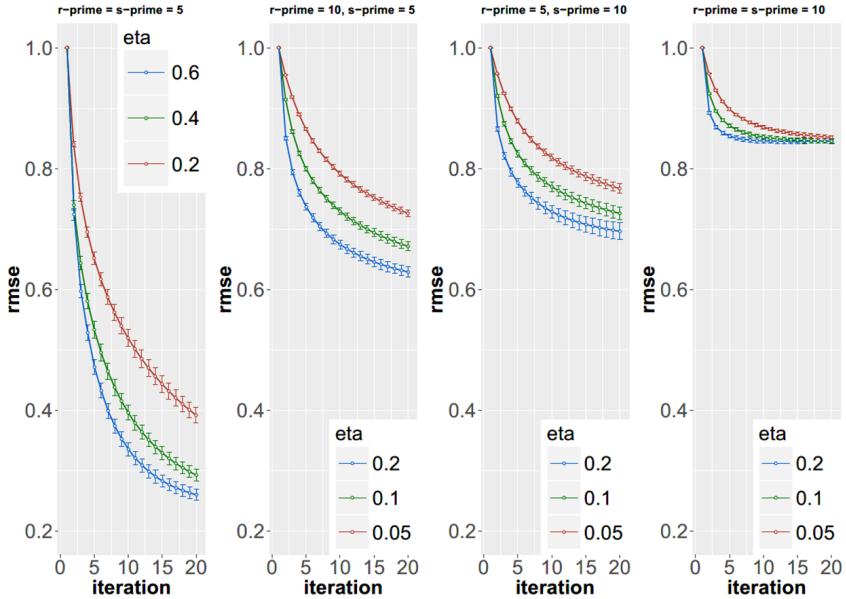

Figure 3: Case 3a: Sparse slices of low-rank matrices

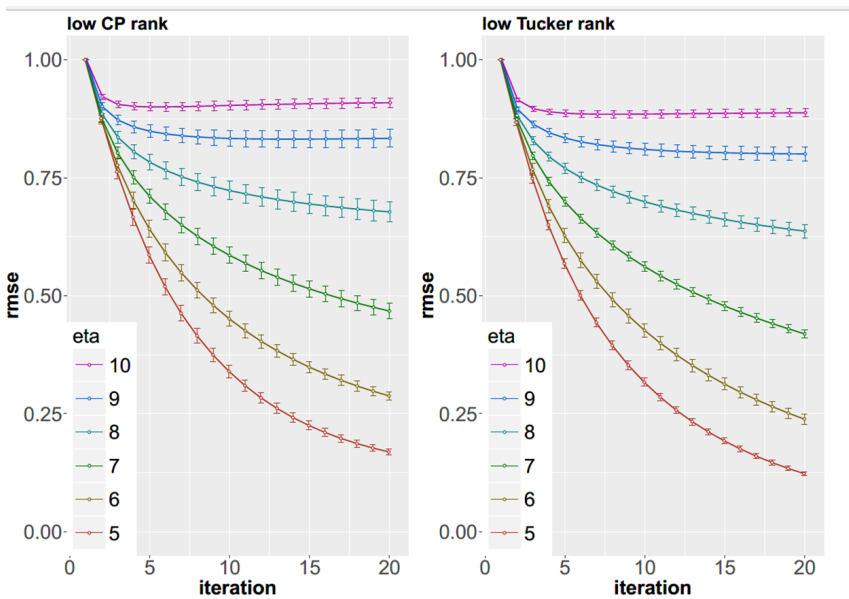

Figure 4: Case 4a, 5a: 4th order tensor

Simulations

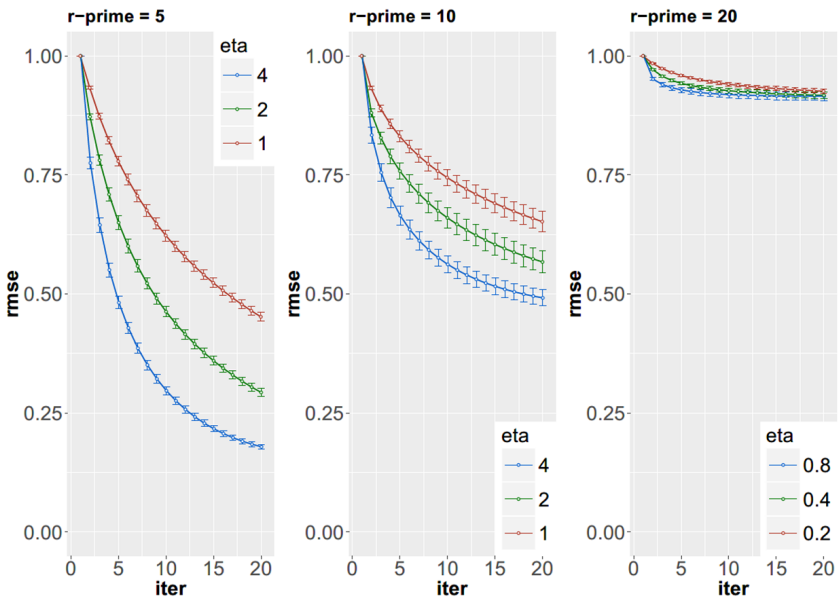

Figure 5: Case 1b: (Logistic) Low CP rank

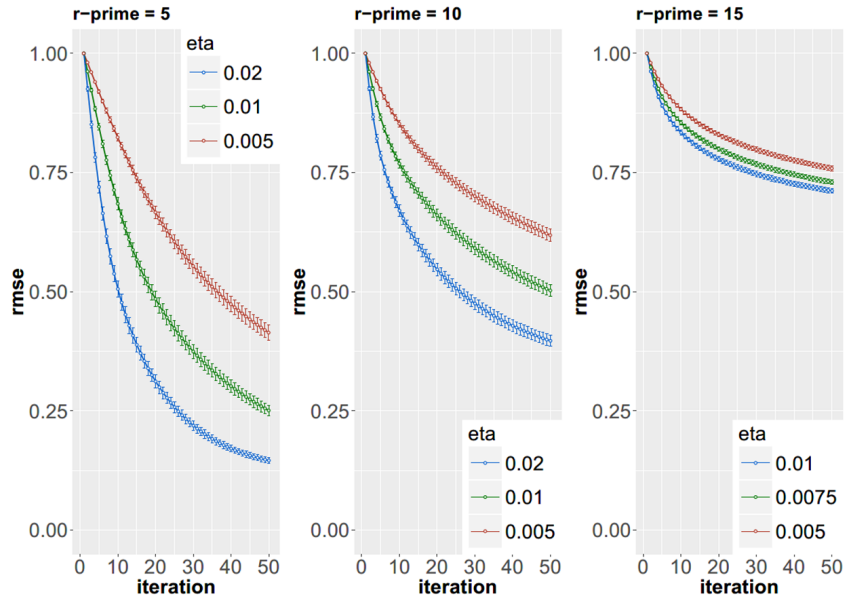

Figure 6: Case 1c: (Poisson) Low CP rank

Simulations

rmse (sd)	SNR	PGD	Convex Regularization
Case 6a	High	**0.11 (0.01)**	0.28 (0.02)
	Moderate	**0.22 (0.01)**	0.47 (0.02)
	Low	**0.46 (0.03)**	0.69 (0.02)
Case 7a	High	**0.07 (0.01)**	0.18 (0.01)
	Moderate	**0.14 (0.01)**	0.32 (0.02)
	Low	**0.28 (0.02)**	0.51 (0.02)
Case 8a	High	**0.08 (0.01)**	0.12 (0.01)
	Moderate	**0.16 (0.01)**	0.23 (0.01)
	Low	**0.30 (0.01)**	0.41 (0.02)
Case 6b	High	**0.16 (0.01)**	0.44 (0.02)
	Moderate	**0.20 (0.01)**	0.54 (0.02)
	Low	**0.35 (0.02)**	0.66 (0.02)
Case 7b	High	**0.17 (0.01)**	0.46 (0.02)
	Moderate	**0.22 (0.01)**	0.55 (0.02)
	Low	**0.35 (0.01)**	0.67 (0.01)
Case 8b	High	**0.26 (0.01)**	0.37 (0.02)
	Moderate	**0.34 (0.02)**	0.50 (0.01)
	Low	**0.56 (0.04)**	0.68 (0.02)
Case 6c	High	**0.09 (0.01)**	0.57 (0.03)
	Moderate	**0.17 (0.01)**	0.61 (0.04)
	Low	**0.39 (0.04)**	0.71 (0.03)
Case 7c	High	**0.12 (0.01)**	0.74 (0.02)
	Moderate	**0.21 (0.02)**	0.75 (0.02)
	Low	**0.43 (0.06)**	0.80 (0.02)
Case 8c	High	**0.13 (0.01)**	0.79 (0.03)
	Moderate	**0.22 (0.03)**	0.81 (0.03)
	Low	**0.32 (0.03)**	0.83 (0.02)

Table 1: rmse of nonconvex PGD vs convex regularization

Simulations

rmse (sd)	SNR	sequential low-rank projection	naive matricization
Case 9a	High	**0.22 (0.01)**	0.57 (0.01)
	Moderate	**0.24 (0.01)**	0.65 (0.01)
	Low	**0.31 (0.01)**	0.80 (0.01)
Case 10a	High	**0.10 (0.01)**	0.58 (0.01)
	Moderate	**0.35 (0.01)**	0.69 (0.01)
	Low	**0.35 (0.01)**	0.85 (0.01)
Case 11a	High	**0.15 (0.01)**	0.56 (0.01)
	Moderate	**0.29 (0.01)**	0.67 (0.01)
	Low	**0.57 (0.02)**	0.83 (0.01)

Table 2: rmse of approximate Tucker projection vs naive matricization