Convergence Rates of Biased Stochastic Optimization for Learning Sparse Ising Models

Jean Honorio JHONORIO@CS.SUNYSB.EDU
Stony Brook University, Stony Brook, NY 11794, USA

Abstract

We study the convergence rate of *stochastic* optimization of *exact* (NP-hard) objectives, for which only biased estimates of the gradient are available. We motivate this problem in the context of learning the structure and parameters of Ising models. We first provide a convergence-rate analysis of *deterministic* errors for *forward-backward splitting* (FBS). We then extend our analysis to *biased stochastic* errors, by first characterizing a family of samplers and providing a high probability bound that allows understanding not only FBS, but also *proximal gradient* (PG) methods. We derive some interesting conclusions: FBS requires only a logarithmically increasing number of random samples in order to converge (although at a very low rate); the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG; accelerated PG is not guaranteed to converge in the biased stochastic setting.

1. Introduction

Structure learning aims to discover the topology of a probabilistic network of variables such that this network represents accurately a given dataset while maintaining low complexity. Accuracy of representation is measured by the likelihood that the model explains the observed data, while complexity of a graphical model is measured by its number of parameters.

One challenge of structure learning is that the number of possible structures is super-exponential in the number of variables. For Ising models, the number of parameters, the number of edges in the structure and the number of non-zero elements in the *ferro-magnetic coupling* matrix are equivalent measures of model complexity. Therefore a computationally tractable approach is to use sparseness promoting regularizers (Wainwright et al., 2006; Banerjee et al., 2008; Höfling & Tibshirani, 2009).

One additional challenge for Ising models (and Markov random fields in general) is that computing the likelihood of a candidate structure is NP-hard. For this reason, several researchers propose exact optimization of approximate objectives, such as ℓ_1-regularized logistic regression (Wainwright et al., 2006), greedy optimization of the conditional log-likelihoods (Jalali et al., 2011), pseudo-likelihood (Besag, 1975) and a sequence of first-order approximations of the exact log-likelihood (Höfling & Tibshirani, 2009). Several convex upper bounds and approximations to the log-partition function have been proposed for maximum likelihood estimation, such as the log-determinant relaxation (Banerjee et al., 2008), the cardinality bound (El Ghaoui & Gueye, 2008), the Bethe entropy (Lee et al., 2006; Parise & Welling, 2006), tree-reweighted approximations and general weighted free-energy (Yang & Ravikumar, 2011).

In this paper, we focus on the stochastic optimization of the exact log-likelihood as our motivating problem. The use of *stochastic maximum likelihood* dates back to (Geyer, 1991; Younes, 1988), in which Markov chain Monte Carlo (MCMC) was used for approximating the gradient. For restricted Boltzmann machines (a very related graphical model) researchers have proposed a variety of approximation methods, such as variational approximations (Murray & Ghahramani, 2004), contrastive divergence (Hinton, 2002), persistent contrastive divergence (Tieleman, 2008), tempered MCMC (Salakhutdinov, 2009; Desjardins et al., 2010), adaptive MCMC (Salakhutdinov, 2010) and particle filtering (Asuncion et al., 2010).

Empirical results in (Marlin et al., 2010) suggests that stochastic maximum likelihood is superior to con-

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

trastive divergence, pseudo-likelihood, ratio matching and generalized score matching for learning restricted Boltzmann machines, in the sense that it produces a higher test set log-likelihood, and more consistent classification results across datasets.

Learning sparse Ising models leads to the use of stochastic optimization with biased estimates of the gradient. Most work in stochastic optimization assumes the availability of unbiased estimates (Duchi & Singer, 2009; Duchi et al., 2010; Hu et al., 2009; Nemirovski et al., 2009). Additionally, other researchers have analyzed convergence rates in the presence of *deterministic* errors that do not decrease over time (d'Aspremont, 2008; Baes, 2009; Devolder et al., 2011) and show convergence up to a constant level. Similarly, Devolder (2012) analyzed the case of *stochastic* errors with *fixed* bias and variance and show convergence up to a constant level.

Notable exceptions are the recent works of Schmidt et al. (2011); Friedlander & Schmidt (2011); Duchi et al. (2011). Schmidt et al. (2011) analyzed *proximal-gradient* (PG) methods for *deterministic* errors of the gradient that decrease over time, for inexact projection steps and Lipschitz as well as strongly convex functions. In our work, we restrict our analysis to exact projection steps and do not assume strong convexity. Both assumptions are natural for learning sparse models under the ℓ_1 regularization. Friedlander & Schmidt (2011) provides convergence rates in expected value for PG with *stochastic* errors that decrease over time in expected value. Friedlander & Schmidt (2011) proposes a growing sample-size strategy for approximating the gradient, i.e. by picking an increasing number of training samples in order to better approximate the gradient. In contrast, our work is for NP-hard gradients and we provide bounds with high probability, by taking into account the bias and the variance of the errors. Duchi et al. (2011) analyzed *mirror descent* (a generalization that includes forward-backward splitting) and show convergence rates in expected value and with high probability with respect to the mixing time of the sampling distribution. We argue that practitioners usually terminate Markov chains before properly mixing, and therefore we motivate our analysis for a controlled increasing number of random samples.

Regarding our contribution in optimization, we provide a convergence-rate analysis of *deterministic* errors for three different flavors of *forward-backward splitting* (FBS): robust (Nemirovski et al., 2009), basic and random (Duchi & Singer, 2009). We extend our analysis to *biased stochastic* errors, by first characterizing a family of samplers (including importance sampling and MCMC) and providing a high probability bound that is useful for understanding the convergence of not only FBS, but also PG (Schmidt et al., 2011). Our analysis shows the bias/variance term and allow to derive some interesting conclusions. First, FBS for deterministic or biased stochastic errors requires only a logarithmically increasing number of random samples in order to converge (although at a very low rate). More interestingly, we found that the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG. We also found that accelerated PG is not guaranteed to converge in the biased stochastic setting.

Regarding our contribution in structure learning, we show that the optimal solution of maximum likelihood estimation is bounded (to the best of our knowledge, this has not been shown before). Our analysis shows provable convergence guarantees for finite iterations and finite number of random samples. Note that while consistency in structure recovery has been established (e.g. Wainwright et al. (2006)), convergence rates of parameter learning for fixed structures is up to now unknown. Our analysis can be easily extended to Markov random fields with higher order cliques as well as parameter learning for fixed structures by using a ℓ_2^2 regularizer instead.

2. Our Motivating Problem

In this section, we introduce the problem of learning sparse Ising models and discuss its properties. Our discussion will motivate a set of bounds and assumptions for a more general convergence rate analysis.

2.1. Problem Setup

An *Ising model* is a Markov random field on binary variables with pairwise interactions. It first arose in statistical physics as a model for the energy of a physical system of interacting atoms (Koller & Friedman, 2009). Formally, the probability mass function (PMF) of an Ising model parameterized by $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$ is defined as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{W}, \mathbf{b})} e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}} \tag{1}$$

where the domain for the binary variables is $\mathbf{x} \in \{-1, +1\}^N$, $\mathbf{W} \in \mathbb{R}^{N \times N}$ is symmetric with zero diagonal, $\mathbf{b} \in \mathbb{R}^N$ and partition function is defined as $\mathcal{Z}(\mathbf{W}, \mathbf{b}) = \sum_{\mathbf{x}} e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}}$. For clarity of the convergence rate analysis, we also define $\boldsymbol{\theta} \in \mathbb{R}^M$ where $M = N^2$.

In the physics literature, $\mathbf{W}$ and $\mathbf{b}$ are called *ferromagnetic coupling* and *external magnetic field* respec-

tively. $\mathbf{W}$ defines the topology of the Markov random field, i.e. the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as $\mathcal{V} = \{1, \ldots, N\}$ and $\mathcal{E} = \{(n_1, n_2) \mid n_1 < n_2 \wedge w_{n_1 n_2} \neq 0\}$. It is well known that, for an Ising model with arbitrary topology, computing the partition function $\mathcal{Z}$ is NP-hard (Barahona, 1982). It is also NP-hard to approximate $\mathcal{Z}$ with high probability and arbitrary precision (Chandrasekaran et al., 2008).

The number of edges $|\mathcal{E}|$ or equivalently the cardinality (number of non-zero entries) of $\mathbf{W}$ is a measure of model complexity, and it can be used as a regularizer for maximum likelihood estimation. The main disadvantage of using such penalty is that it leads to a NP-hard problem, regardless of the computational complexity of the log-likelihood.

Next, we formalize the problem of finding a sparse Ising model by regularized maximum likelihood estimation. We replace the cardinality penalty by the ℓ_1-norm regularizer as in (Wainwright et al., 2006; Banerjee et al., 2008; Höfling & Tibshirani, 2009).

Given a complete dataset with T i.i.d. samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$, and a sparseness parameter $\rho > 0$ the ℓ_1-regularized maximum likelihood estimation for the Ising model in eq.(1) becomes:

$$\min_{\mathbf{W},\mathbf{b}} \mathcal{L}(\mathbf{W}, \mathbf{b}) + \mathcal{R}(\mathbf{W}) \tag{2}$$

where the negative (average) log-likelihood $\mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{T}\sum_t \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}) = \log \mathcal{Z}(\mathbf{W}, \mathbf{b}) - \langle \widehat{\boldsymbol{\Sigma}}, \mathbf{W} \rangle - \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{b}$, the empirical second-order moment $\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_t \mathbf{x}^{(t)}{\mathbf{x}^{(t)}}^{\mathrm{T}} - \mathbf{I}$, the empirical first-order moment $\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_t \mathbf{x}^{(t)}$ and the regularizer $\mathcal{R}(\mathbf{W}) = \rho\|\mathbf{W}\|_1$.

The objective function in eq.(2) is convex, given the convexity of the log-partition function (Koller & Friedman, 2009), linearity of the scalar products and convexity of the non-smooth ℓ_1-norm regularizer. As discussed before, computing the partition function $\mathcal{Z}$ is NP-hard, and so is computing the objective function in eq.(2).

2.2. Bounds

In what follows, we show boundedness of the optimal solution and the gradients of the maximum likelihood problem. Both are important ingredients for showing convergence and are largely used assumptions in optimization. In this paper, we follow the original formulation of the problem given in (Wainwright et al., 2006; Banerjee et al., 2008; Höfling & Tibshirani, 2009), which does not regularize $\mathbf{b}$. We found interesting to show that this problem has bounds for $\|\mathbf{b}^*\|_1$ unlike other stochastic optimization problems, e.g. SVMs (Shalev-Shwartz et al., 2007).

First, we make some observations that will help us derive our bounds. The empirical second-order moment $\widehat{\boldsymbol{\Sigma}}$ and first-order moment $\widehat{\boldsymbol{\mu}}$ in eq.(2) are computed from binary variables in $\{-1, +1\}$, therefore $\|\widehat{\boldsymbol{\Sigma}}\|_\infty \leq 1$ and $\|\widehat{\boldsymbol{\mu}}\|_\infty \leq 1$.

Assumption 1. *It is reasonable to assume that the empirical first-order moment of every variable is not equal to -1 (or $+1$), since this would be equivalent to observe a constant value -1 (or $+1$) for such variables in every sample in the dataset, i.e.* $(\exists n)\ |\widehat{\mu}_n| = 1 \Leftrightarrow (\forall t)\ x_n^{(t)} = -1 \vee (\forall t)\ x_n^{(t)} = 1$. *Therefore, we assume* $\|\widehat{\boldsymbol{\mu}}\|_\infty < 1 \Leftrightarrow (\forall n) -1 < \widehat{\mu}_n < +1$.

Given those observations, we state our bounds in the following theorem. For clarity of the convergence rate analysis, we also define the bound D of the optimal solution.

Theorem 2. *The optimal solution $\boldsymbol{\theta}^* = (\mathbf{W}^*, \mathbf{b}^*)$ of the maximum likelihood problem in eq.*(2) *is bounded as follows:*

$$\begin{array}{lll} \text{i.} & \|\mathbf{W}^*\|_1 & \leq \frac{N \log 2}{\rho} \\ \text{ii.} & \|\mathbf{b}^*\|_1 & \leq \frac{N \log 2(\rho + 1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)}{\rho(1 - \|\widehat{\boldsymbol{\mu}}\|_\infty)} \\ \text{iii.} & \|\boldsymbol{\theta}^*\|_2 & \leq D \end{array} \tag{3}$$

where $D^2 = \left(\frac{N \log 2}{\rho}\right)^2 \left(1 + \left(\frac{\rho + 1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty}{1 - \|\widehat{\boldsymbol{\mu}}\|_\infty}\right)^2\right)$.

Proof Sketch. Claim i and ii follow from the fact that the function evaluated at $(\mathbf{W}^*, \mathbf{b}^*)$ is less than at $(\mathbf{0}, \mathbf{0})$. Additionally, Claim i follows from non-negativity of the negative log-likelihood in eq.(2), while Claim ii follows from non-negativity of the regularizer and from Assumption 1. Claim iii follows from norm inequalities and Claims i and ii. □

(Please, see Appendix C for detailed proofs.)

If we choose to add the regularizer $\rho\|\mathbf{b}\|_1$ in eq.(2), it is easy to conclude that $\|\mathbf{W}^*\|_1 + \|\mathbf{b}^*\|_1 \leq \frac{N \log 2}{\rho}$ as in Claim i of Theorem 2.

The gradient of the objective function of the maximum likelihood problem in eq.(2) is defined as:

$$\begin{array}{lll} \text{i.} & \partial \log \mathcal{Z}/\partial \mathbf{W} & = \mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] \\ \text{ii.} & \partial \log \mathcal{Z}/\partial \mathbf{b} & = \mathbb{E}_{\mathcal{P}}[\mathbf{x}] \\ \text{iii.} & \partial \mathcal{L}/\partial \mathbf{W} & = \partial \log \mathcal{Z}/\partial \mathbf{W} - \widehat{\boldsymbol{\Sigma}} \\ \text{iv.} & \partial \mathcal{L}/\partial \mathbf{b} & = \partial \log \mathcal{Z}/\partial \mathbf{b} - \widehat{\boldsymbol{\mu}} \end{array} \tag{4}$$

where $\mathcal{P}$ is the probability distribution with PMF $p_{\boldsymbol{\theta}}(\mathbf{x})$. The expression in eq.(4) uses the fact that $\mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^{\mathrm{T}} p_{\boldsymbol{\theta}}(\mathbf{x})$ and $\mathbb{E}_{\mathcal{P}}[\mathbf{x}] = \sum_{\mathbf{x}} \mathbf{x} p_{\boldsymbol{\theta}}(\mathbf{x})$.

It is well known that computing the gradients $\partial \log \mathcal{Z}/\partial \mathbf{W}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b}$ is NP-hard. The complexity results in (Chandrasekaran et al., 2008) imply that approximating those gradients with high probability and arbitrary precision is also NP-hard.

Next, we state some properties of the gradient of the exact log-likelihood. For clarity of the convergence rate analysis, we also define the Lipschitz constant G.

Lemma 3. *The objective function of the maximum likelihood problem in eq.*(2) *has the following Lipschitz continuity properties:*

$$\begin{array}{ll} \text{i.} & \|\partial \log \mathcal{Z}/\partial \mathbf{W}\|_\infty \ , \ \|\partial \log \mathcal{Z}/\partial \mathbf{b}\|_\infty \leq 1 \\ \text{ii.} & \|\partial \mathcal{L}/\partial \mathbf{W}\|_\infty \ \leq 1 + \|\widehat{\mathbf{\Sigma}}\|_\infty \\ \text{iii.} & \|\partial \mathcal{L}/\partial \mathbf{b}\|_\infty \ \ \leq 1 + \|\widehat{\boldsymbol{\mu}}\|_\infty \\ \text{iv.} & \|\partial \mathcal{R}/\partial \mathbf{W}\|_\infty \leq \rho \\ \text{v.} & \|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2 \ , \ \|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 \leq G \end{array} \tag{5}$$

where $G^2 = N^2 \max((1+\|\widehat{\mathbf{\Sigma}}\|_\infty)^2 + \frac{1}{N}(1+\|\widehat{\boldsymbol{\mu}}\|_\infty)^2, \rho^2)$.

Proof Sketch. Claims i to iii follow from the fact that the terms $\partial \log \mathcal{Z}/\partial \mathbf{W}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b}$ in eq.(4) are the second and first-order moment of binary variables in $\{-1, +1\}$. Claim iv follows from the definition of subgradients. Claim v follows from norm inequalities and Claims ii to iv. □

2.3. Approximating the Gradient of the Log-Partition Function

Suppose one wants to evaluate the expression $\mathbb{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ in eq.(4) which is the gradient of the log-partition function. Let assume we know the distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ up to a constant factor, i.e. $p'_{\boldsymbol{\theta}}(\mathbf{x}) = e^{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}}$. Importance sampling draws S samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ from a trial distribution with PMF $q(\mathbf{x})$, calculates the importance weights $\alpha^{(s)} = p'_{\boldsymbol{\theta}}(\mathbf{x}^{(s)})/q(\mathbf{x}^{(s)})$ and produces the estimate $(\sum_s \alpha^{(s)} \mathbf{x}^{(s)} \mathbf{x}^{(s)^{\mathrm{T}}})/\sum_s \alpha^{(s)}$. On the other hand, MCMC generates S samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ from the distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ based on constructing a Markov chain whose stationary distribution is $p_{\boldsymbol{\theta}}(\mathbf{x})$. Thus, the estimate becomes $\frac{1}{S}\sum_s \mathbf{x}^{(s)} \mathbf{x}^{(s)^{\mathrm{T}}}$.

In what follows, we characterize a family of samplers that includes importance sampling and MCMC as shown in (Peskun, 1973; Liu, 2001).

Definition 4. *A* (B, V, S, D)*-sampler takes* S *random samples from a distribution* $\mathcal{Q}$ *and produces biased estimates of the gradient of the log-partition function* $\partial \log \mathcal{Z}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}$*, with error* $\boldsymbol{\xi}$ *that has bias and variance:*

$$\begin{array}{ll} \text{i.} & \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}\|_2] \ \ \leq \frac{B}{S} + \mathcal{O}(\frac{1}{S^2}) \\ \text{ii.} & \mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}\|_2] \leq \frac{V}{S} + \mathcal{O}(\frac{1}{S^2}) \end{array} \tag{6}$$

for $B \geq 0, V \geq 0$ *and* $(\forall \boldsymbol{\theta})\ \|\boldsymbol{\theta}\|_2 \leq D$.

Note that a (B, V, S, D)-sampler is asymptotically unbiased with asymptotically vanishing variance, i.e. $S \to +\infty \Rightarrow \frac{B}{S} \to 0 \wedge \frac{V}{S} \to 0$. Unfortunately, analytical approximations of the constants B and V are difficult to obtain even for specific classes, e.g. Ising models. The theoretical analysis implies that such constants B and V exist (Peskun, 1973; Liu, 2001) for importance sampling and MCMC. We argue that this apparent disadvantage does not diminish the relevance of our analysis, since we can reasonably expect that more refined samplers lead to lower B and V.

Note that Definition 4 does not contradict the complexity results in (Chandrasekaran et al., 2008) that show that it is likely impossible to approximate $\mathcal{Z}$ (and therefore its gradient) with probability greater than $1 - \delta$ and arbitrary precision ε in time polynomial in $\log \frac{1}{\delta}$ and $\frac{1}{\varepsilon}$. Definition 4 assumes biasedness and a polynomial decay instead of an exponential decay (which is a more stringent condition) and cannot be used to derive two-sided high probability bounds that are both $\mathcal{O}(\log \frac{1}{\delta})$ and $\mathcal{O}(\frac{1}{S})$. Therefore, Definition 4 cannot be used to obtain polynomial-time algorithms as the ones considered in (Chandrasekaran et al., 2008).

Assumption 5. *It is reasonable to assume that the estimates of the gradient of the log-partition function are inside* $[-1; +1]$ *since they are approximations of the second and first-order moment of binary variables in* $\{-1, +1\}$*. Furthermore, it is straightforward to enforce Lipschitz continuity (condition i of Lemma 3) for any sampler (e.g. importance sampling, MCMC or any conceivable method) by limiting its output to be inside* $[-1; +1]$*. More formally, we have:*

$$\begin{array}{ll} \text{i.} & \|\partial \log \mathcal{Z}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}\|_\infty \leq 1 \\ \text{ii.} & \|\partial \mathcal{L}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}\|_2 \ \ \ \ \ \ \leq G \end{array} \tag{7}$$

3. Biased Stochastic Optimization

In this section, we analyze the convergence rates of *forward-backward splitting*. Our results apply to any problem that fulfills the following largely used assumptions in optimization:

- the objective function is composed by a smooth function $\mathcal{L}(\boldsymbol{\theta})$ and non-smooth regularizer $\mathcal{R}(\boldsymbol{\theta})$
- the optimal solution is bounded, i.e. $\|\boldsymbol{\theta}^*\|_2 \leq D$
- each visited point is at a bounded distance from the optimal solution, i.e. $(\forall k)\ \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2 \leq D$
- both $\mathcal{L}$ and $\mathcal{R}$ are Lipschitz continuous, i.e. $\|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2 \ , \ \|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 \leq G$
- the non-smooth regularizer vanishes at zero, i.e. $\mathcal{R}(\mathbf{0}) = 0$

We additionally require that the errors do not change the Lipschitz continuity properties, i.e. $\|\partial\mathcal{L}/\partial\boldsymbol{\theta} + \boldsymbol{\xi}\|_2 \leq G$ (as discussed in Assumption 5).

3.1. Algorithm

We analyze *forward-backward splitting* (Duchi & Singer, 2009) for deterministic as well as biased stochastic errors, for non-increasing step sizes of the form $\eta_k \in \mathcal{O}(\frac{1}{k^r})$ for $0 < r < 1$. This method is equivalent to basic *proximal gradient* (Schmidt et al., 2011) for $r = 0$ (constant step size). We point out that FBS has $\mathcal{O}(\frac{1}{\sqrt{K}})$ convergence for $r = \frac{1}{2}$, while basic PG has $\mathcal{O}(\frac{1}{K})$ convergence, and accelerated PG has $\mathcal{O}(\frac{1}{K^2})$ convergence. Thus, PG methods have faster convergence but they are more sensitive to errors.

FBS performs gradient descent steps for the smooth part of the objective function, and (closed form) projection steps for the non-smooth part. Here we assume that at each iteration k, we approximate the gradient with some (deterministic or biased stochastic) error $\boldsymbol{\xi}^{(k)}$. For our objective function in eq.(2), one iteration of the algorithm is equivalent to:

$$\begin{aligned} &\text{i. } \boldsymbol{\theta}^{(k+\frac{1}{2})} = \boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_{\mathcal{L}}^{(k)} + \boldsymbol{\xi}^{(k)}) \\ &\text{ii. } \boldsymbol{\theta}^{(k+1)} = \arg\min_{\boldsymbol{\theta}}(\tfrac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k+\frac{1}{2})}\|_2^2 + \eta_{k+1}\mathcal{R}(\boldsymbol{\theta})) \end{aligned} \tag{8}$$

where $\mathbf{g}_{\mathcal{L}}^{(k)} = \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^{(k)})$, and $\boldsymbol{\xi}^{(k)}$ is the error in the gradient approximation. Step ii is a projection step for the non-smooth regularizer $\mathcal{R}(\boldsymbol{\theta})$. For the regularizer in our motivating problem $\mathcal{R}(\mathbf{W}) = \rho\|\mathbf{W}\|_1$, Step ii decomposes into N^2 independent *lasso* problems.

3.2. Convergence Rates for Deterministic Errors

In what follows, we analyze three different flavors of forward-backward splitting: *robust* which outputs the weighted average of all visited points by using the step sizes as in *robust stochastic approximation* (Nemirovski et al., 2009), *basic* which outputs the average of all visited points as in (Duchi & Singer, 2009), or *random* which outputs a point chosen uniformly at random from the visited points. Here we assume that at each iteration k, we approximate the gradient with some deterministic error $\boldsymbol{\xi}^{(k)}$. Our results in this subsection will allow us to draw some conclusions regarding not only FBS but also proximal gradient.

In order to make our bounds more general for different choices of step size $\eta_k \in \mathcal{O}(\frac{1}{k^r})$ for some $0 < r < 1$, we use *generalized harmonic numbers* $H_{r,K} = \sum_{k=1}^{K} \frac{1}{k^r}$ and therefore $H_{0,K} = K$, $H_{r,K} \approx \frac{K^{1-r}}{1-r}$ for $0 < r < 1$, $H_{1,K} \approx \log K$ and $H_{r,K} \approx \frac{1-K^{1-r}}{r-1}$ for $r > 1$.

Additionally, we define a weighted error term that will be used for our analysis of deterministic as well as biased stochastic errors. Given a sequence of errors $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$ and a set of arbitrary weights γ_k such that $\sum_k \gamma_k = 1$, the error term is defined as:

$$A_{\boldsymbol{\gamma},\boldsymbol{\xi}} \equiv \textstyle\sum_k \gamma_k \|\boldsymbol{\xi}^{(k)}\|_2 \tag{9}$$

First, we show the convergence rate of robust FBS.

Theorem 6. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$, step size $\eta_k = \frac{\beta}{Gk^r}$ for $0 < r < 1$, initial point $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, the objective function evaluated at the weighted average of all visited points converges to the optimal solution with rate:*

$$\begin{aligned} &\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \pi_\eta(K) \\ &\leq \tfrac{D^2 G}{2\beta H_{r,K}} + 2DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \tfrac{4\beta G H_{2r,K}}{H_{r,K}} \end{aligned} \tag{10}$$

where $\overline{\boldsymbol{\theta}} = \frac{\sum_k \eta_k \boldsymbol{\theta}^{(k)}}{\sum_k \eta_k}$, the weighted average regret $\pi_\eta(K) = \frac{\sum_k \eta_k(\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))}{\sum_k \eta_k} - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*)$, the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is defined as in eq.(9), and the error weights $\gamma_k = \frac{1/k^r}{H_{r,K}}$ such that $\sum_k \gamma_k = 1$.

Proof Sketch. By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k \eta_k(\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))/\sum_k \eta_k$. Then we apply a technical lemma for bounding consecutive steps (Please, see Appendix B). □

Second, we show the convergence rate of basic FBS.

Theorem 7. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$, step size $\eta_k = \frac{\beta}{Gk^r}$ for $0 < r < 1$, initial point $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, the objective function evaluated at the average of all visited points converges to the optimal solution with rate:*

$$\begin{aligned} &\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \pi(K) \\ &\leq \tfrac{D^2 G(K+1)^r}{2\beta K} + 2^{1+r} DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \tfrac{2^{2+r}\beta G H_{r,K}}{K} \end{aligned} \tag{11}$$

where $\overline{\boldsymbol{\theta}} = \frac{\sum_k \boldsymbol{\theta}^{(k)}}{K}$, the average regret $\pi(K) = \frac{\sum_k (\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))}{K} - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*)$, the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is defined as in eq.(9), and the error weights $\gamma_k = \frac{1}{K}$ such that $\sum_k \gamma_k = 1$.

Proof Sketch. By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k (\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}))/K$. Then we apply a technical lemma for bounding consecutive steps (Please, see Appendix B). □

Finally, we show the convergence rate of random FBS.

Theorem 8. *For a sequence of deterministic errors* $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$*, step size* $\eta_k = \frac{\beta}{Gk^r}$ *for* $0 < r < 1$*, initial point* $\boldsymbol{\theta}^{(1)} = \mathbf{0}$ *and some confidence parameter* $0 < \varepsilon < 1$*, the objective function evaluated at a point* k *chosen uniformly at random from the visited points converges, with probability at least* $1 - \varepsilon$*, to the optimal solution with rate:*

$$\mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathcal{R}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \frac{1}{\varepsilon}\left(\frac{D^2 G (K+1)^r}{2\beta K} + 2^{1+r} D A_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \frac{2^{2+r}\beta G H_{r,K}}{K}\right) \quad (12)$$

where the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ *is defined as in eq.*(9)*, and the error weights* $\gamma_k = \frac{1}{K}$ *such that* $\sum_k \gamma_k = 1$*.*

Proof Sketch. Since the distribution is uniform on k, the expected value of the objective function is equal to the average of the objective function evaluated at all visited points, i.e. the average regret $\pi(K)$. The final result follows from Markov's inequality and the upper bound of $\pi(K)$ given in Theorem 7. □

The convergence rates in Theorems 6, 7 and 8 lead to an error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ that is linear, while the error term is quadratic in the analysis of proximal gradient (Schmidt et al., 2011). In basic PG, the error term can be written as:

$$\tfrac{1}{K}\left(\textstyle\sum_k \|\boldsymbol{\xi}^{(k)}\|_2\right)^2 = K(A_{\boldsymbol{\gamma},\boldsymbol{\xi}})^2 \quad (13)$$

where the error weights $\gamma_k = \frac{1}{K}$ such that $\sum_k \gamma_k = 1$. In accelerated PG, the error term can be written as:

$$\tfrac{4}{(K+1)^2}\left(\textstyle\sum_k k\|\boldsymbol{\xi}^{(k)}\|_2\right)^2 = K^2(A_{\boldsymbol{\gamma},\boldsymbol{\xi}})^2 \quad (14)$$

where the error weights $\gamma_k = k/\binom{K}{2}$ so that $\sum_k \gamma_k = 1$.

Note that both PG methods contain terms K and K^2, which are not in our analysis. As noted in (Schmidt et al., 2011), errors have a greater effect on the accelerated method than on the basic method. This observation suggests that, unlike in the error-free case, accelerated PG is not necessarily better than the basic method due to a higher sensitivity to errors (Devolder et al., 2011).

Intuitively speaking, basic PG is similar to basic FBS in the sense that errors from all iterations have the same effect on the convergence rate, i.e. γ_k is constant. In robust FBS, errors in the last iterations have a lower effect on the convergence rate than errors in the beginning, i.e. γ_k is decreasing. In accelerated PG, errors in the last iterations have a bigger effect on the convergence rate than errors in the beginning, i.e. γ_k is increasing.

The analysis of Schmidt et al. (2011) for deterministic errors implies that in order to have convergence, the errors must decrease at a rate $\|\boldsymbol{\xi}^{(k)}\|_2 \in \mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$ for some $\epsilon > 0$ in the case of basic PG, and $\mathcal{O}(\frac{1}{k^{1+\epsilon}})$ for accelerated PG. In contrast, our analysis of FBS show that we only need logarithmically decreasing errors $\mathcal{O}(\frac{1}{\log k})$ in order to have convergence. Regarding $\mathcal{O}(\frac{1}{\sqrt{K}})$ convergence of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$, basic and robust FBS requires errors $\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$ (the minimum required for convergence in basic PG). Table 1 summarizes the requirements for different convergence rates of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ of FBS as well as the error terms of basic PG in eq.(13) and accelerated PG in eq.(14).

Table 1. Order of errors $\|\boldsymbol{\xi}^{(k)}\|_2$ required to obtain convergence of the error term for the *deterministic* case: basic (PB) and accelerated (PA) proximal gradient, basic (FB) and robust (FR) forward-backward splitting.

Method	**Convergence**			
	for $K \to +\infty$	$\mathcal{O}(\frac{1}{\sqrt{K}})$	$\mathcal{O}(\frac{1}{K})$	$\mathcal{O}(\frac{1}{K^2})$
PB	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{3/4+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	-
PA	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{5/4+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{3/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{2+\epsilon}})$
FB ($r=\frac{1}{2}$)	$\mathcal{O}(\frac{1}{\log k})$	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	$\mathcal{O}(\frac{1}{k^{1+\epsilon}})$	-
FR ($r=\frac{1}{2}$)	$\mathcal{O}(\frac{1}{\log k})$	$\mathcal{O}(\frac{1}{k^{1/2+\epsilon}})$	-	-

For an informal (and incomplete) analysis of the results in (Schmidt et al., 2011) for biased stochastic optimization, consider each error bounded by its bias and variance $\|\boldsymbol{\xi}^{(k)}\|_2 \leq B/S_k + c\sqrt{V/S_k}$ for some $c > 0$ and an increasing number of random samples S_k that allows to obtain decreasing errors. Without noting the possible need of "uniform convergence" of the bound for all K iterations (making c a function of K), the number of random samples must increase (at least) at a rate that is quadratic of the rate of the errors. For instance, in order to have $\mathcal{O}(\frac{1}{K})$ convergence, basic PG requires errors to be $\mathcal{O}(\frac{1}{k^{1+\epsilon}})$ and therefore it would require (at least) an increasing number of random samples $S_k \in \mathcal{O}(k^{2+\epsilon})$ for some $\epsilon > 0$. Accelerated PG would require (at least) $S_k \in \mathcal{O}(k^{4+\epsilon})$ in order to obtain $\mathcal{O}(\frac{1}{K^2})$ convergence. If we include the fact that c is a function of K, then the required number of random samples would be "worse than quadratic" of the required rate of the errors. Fortunately, a formal analysis in the next subsection shows that this is not the case for all methods except accelerated PG.

3.3. Bounding the Error Term for Biased Stochastic Optimization

In what follows, we focus in the analysis of stochastic errors in order to see if better convergence rates can be obtained than the ones informally outlined in the previous subsection. A formal analysis of the er-

ror terms show that *forward-backward splitting* for biased stochastic errors requires only a logarithmically increasing number of random samples in order to converge, i.e. $S_k \in \mathcal{O}(\log k)$. More interestingly, we found that the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG. On the negative side, we found that accelerated PG is not guaranteed to converge in the biased stochastic setting.

Next, we present our high probability bound for the error term for biased stochastic optimization. One way to bound the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ would be to rely on "uniform convergence" arguments, i.e. to bound the error of each iteration $\|\boldsymbol{\xi}^{(k)}\|_2$ and then use the well-known union bound. We chose to bound the error term itself, by using the fact that errors become independent (but not identically distributed) conditioned to the parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$. We also allow for a different number of random samples S_k for each iteration k.

Theorem 9. *Given K (B, V, S_k, D)-samplers each producing estimates with an error $\boldsymbol{\xi}^{(k)}$, and given a set of arbitrary weights γ_k such that $\sum_k \gamma_k = 1$. For some confidence parameter $0 < \delta < 1$, with probability at least $1 - \delta$, the error term is bounded as follows:*

$$A_{\boldsymbol{\gamma},\boldsymbol{\xi}} \leq \lambda_1 + \tfrac{2\sqrt{M}}{3K} \log \tfrac{1}{\delta} + \sqrt{2\lambda_2 \log \tfrac{1}{\delta} + \tfrac{4M}{9K^2} \log^2 \tfrac{1}{\delta}} \tag{15}$$

where the bias term $\lambda_1 = \min(2\sqrt{M}, B \sum_k \frac{\gamma_k}{S_k})$ and the variance term $\lambda_2 = \min(4M, V \sum_k \frac{\gamma_k^2}{S_k})$.

Proof Sketch. The bias and variance for each $\|\boldsymbol{\xi}^{(k)}\|_2$ are bounded by $\frac{B}{S_k}$ and $\frac{V}{S_k}$ by Definition 4. By Lemma 3 and Assumption 5 we have $\|\boldsymbol{\xi}^{(k)}\|_2 \leq 2\sqrt{M}$ which is the maximum bias, and its square is the maximum variance. By the definition of marginal distribution, we make $\|\boldsymbol{\xi}^{(1)}\|_2, \ldots, \|\boldsymbol{\xi}^{(K)}\|_2$ independent (but not identically distributed) conditioned to the parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$. We then invoke Bernstein inequality for properly defined variables such that it applies to the weighted average $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$. □

It is interesting to note what happens for a fixed number of random samples $S_k \in \mathcal{O}(1)$. In this case, the bias term $\lambda_1 \in \mathcal{O}(1)$ and therefore FBS will not converge. For robust FBS, the variance term $\lambda_2 \in \mathcal{O}(H_{2r,K}/(H_{r,K})^2)$ which for instance for $r = \frac{1}{2}$ we have $\lambda_2 \in \mathcal{O}(\frac{\log K}{K})$. For basic FBS, the variance term $\lambda_2 \in \mathcal{O}(\frac{1}{K})$. Therefore, for the constant number of random samples, the lack of convergence of FBS is explained only by the bias of the sampler and not its variance.

Table 2. Random samples S_k required to obtain convergence of the error term for the *biased stochastic* case: basic (PB) and accelerated (PA) proximal gradient, basic (FB) and robust (FR) forward-backward splitting.

Method	**Convergence**			
	for $K \to +\infty$	$\mathcal{O}(\frac{1}{\sqrt{K}})$	$\mathcal{O}(\frac{1}{K})$	$\mathcal{O}(\frac{1}{K^2})$
PB	$\mathcal{O}(k^{1/2+\epsilon})$	$\mathcal{O}(k^{3/4+\epsilon})$	$\mathcal{O}(k^{1+\epsilon})$	-
PA	-	-	-	-
FB ($r=\frac{1}{2}$)	$\mathcal{O}(\log k)$	$\mathcal{O}(k^{1/2+\epsilon})$	$\mathcal{O}(k^{1+\epsilon})$	-
FR ($r=\frac{1}{2}$)	$\mathcal{O}(\log k)$	$\mathcal{O}(k^{1/2+\epsilon})$	-	-

Table 2 summarizes the requirements for different convergence rates of the error term $A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ of FBS as well as the error terms of basic PG in eq.(13) and accelerated PG in eq.(14). Note that convergence for FBS is guaranteed for a logarithmically increasing number of random samples $S_k \in \mathcal{O}(\log k)$. Moreover, in order to obtain convergence rates of $\mathcal{O}(\frac{1}{\sqrt{K}})$ and $\mathcal{O}(\frac{1}{K})$, the required number of random samples is just the inverse of the required rate of the errors for the deterministic case, and not "worse than quadratic" as outlined in our informal analysis of the previous subsection.

One important conclusion from Theorem 9 is that the upper bound of the error term is $\Omega(\frac{1}{K})$ independently of the bias term λ_1 and the variance term λ_2. This implies that the error term is $\mathcal{O}(\frac{1}{K})$ for any setting of error weights γ_k and number of random samples S_k. The main implication is that the error term in accelerated PG in eq.(14) is constant and therefore the accelerated method is not guaranteed to converge.

4. Experimental Results

We illustrate our theoretical findings with a small synthetic experiment ($N = 15$ variables) since we want to report the log-likelihood at each iteration. We performed 10 repetitions. For each repetition, we generate edges in the ground truth model $\mathbf{W}_g$ with a 50% density. The weight of each edge is generated uniformly at random from $[-1; +1]$. We set $\mathbf{b}_g = \mathbf{0}$. We finally generate a dataset of 50 samples. We used a "Gibbs sampler" by first finding the mean field distribution and then performing 5 Gibbs iterations. We used a step size factor $\beta = 1$ and regularization parameter $\rho = 1/16$. We also include a two-step algorithm, by first learning the structure by ℓ_1-regularized logistic regression (Wainwright et al., 2006) and then learning the parameters by using FBS with belief propagation for gradient approximation. We summarize our results in Figure 1.

Our experiments suggest that stochastic optimiza-

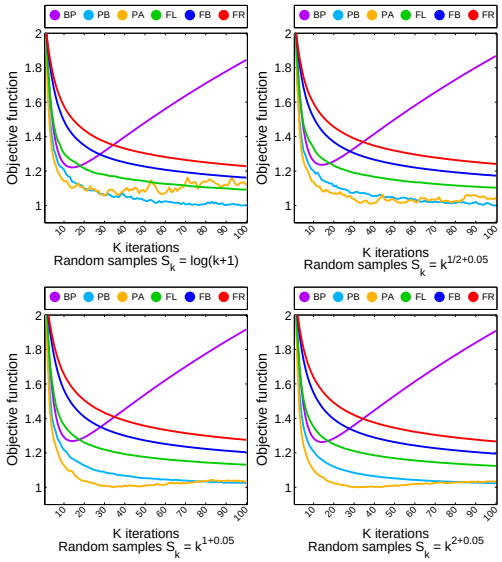

Figure 1. Objective function for different settings of increasing number of random samples. Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

tion converges to the maximum likelihood estimate. We also show the Kullback-Leibler divergence to the ground truth, and more pronounced effects for importance sampling (Please, see Appendix D).

Concluding Remarks. Although we focused on Ising models, the ideas developed in the current paper could be applied to Markov random fields with higher order cliques. Our analysis can be easily extended to parameter learning for fixed structures by using a ℓ_2^2 regularizer instead. Although we show that accelerated proximal gradient is not guaranteed to converge in our specific biased stochastic setting, necessary conditions for its convergence needs to be investigated.

Acknowledgments. This work was done while the author was supported in part by NIH Grants 1 R01 DA020949 and 1 R01 EB007530.

References

Asuncion, A., Liu, Q., Ihler, A., and Smyth, P. Particle filtered MCMC-MLE with connections to contrastive divergence. *ICML*, 2010.

Baes, M. Estimate sequence methods: extensions and approximations. *IFOR internal report, ETH Zurich*, 2009.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 2008.

Barahona, F. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical, Nuclear and General*, 1982.

Besag, J. Statistical analysis of non-lattice data. *The Statistician*, 1975.

Chandrasekaran, V., Srebro, N., and Harsha, P. Complexity of inference in graphical models. *UAI*, 2008.

d'Aspremont, A. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 2008.

Desjardins, G., Courville, A., Bengio, Y., Vincent, P., and Delalleau, O. Parallel tempering for training of restricted Boltzmann machines. *AISTATS*, 2010.

Devolder, O. Stochastic first order methods in smooth convex optimization. *CORE Discussion Papers 2012/9*, 2012.

Devolder, O., Glineur, F., and Nesterov, Y. First-order methods of smooth convex optimization with inexact oracle. *CORE Discussion Papers 2011/2*, 2011.

Duchi, J. and Singer, Y. Efficient online and batch learning using forward backward splitting. *JMLR*, 2009.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. *COLT*, 2010.

Duchi, J., Agarwal, A., Johansson, M., and Jordan, M. Ergodic subgradient descent. *Allerton Conference*, 2011.

El Ghaoui, L. and Gueye, A. A convex upper bound on the log-partition function for binary graphical models. *NIPS*, 2008.

Friedlander, M. and Schmidt, M. Hybrid deterministic-stochastic methods for data fitting. *arXiv:1104.2373*, 2011.

Geyer, C. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics*, 1991.

Hinton, G. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

Höfling, H. and Tibshirani, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *JMLR*, 2009.

Hu, C., Kowk, J., and Pan, W. Accelerated gradient methods for stochastic optimization and online learning. *NIPS*, 2009.

Jalali, A., Johnson, C., and Ravikumar, P. On learning discrete graphical models using greedy methods. *NIPS*, 2011.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Lee, S., Ganapathi, V., and Koller, D. Efficient structure learning of Markov networks using ℓ_1-regularization. *NIPS*, 2006.

Liu, J. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Marlin, B., Swersky, K., Chen, B., and de Freitas, N. Inductive principles for restricted Boltzmann machine learning. *AISTATS*, 2010.

Murray, I. and Ghahramani, Z. Bayesian learning in undirected graphical models: Approximate MCMC algorithms. *UAI*, 2004.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.

Parise, S. and Welling, M. Structure learning in Markov random fields. *NIPS*, 2006.

Peskun, P. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 1973.

Salakhutdinov, R. Learning in Markov random fields using tempered transitions. *NIPS*, 2009.

Salakhutdinov, R. Learning deep Boltzmann machines using adaptive MCMC. *ICML*, 2010.

Schmidt, M., Le Roux, N., and Bach, F. Convergence rates of inexact proximal-gradient methods for convex optimization. *NIPS*, 2011.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal estimated sub-gradient solver for SVM. *ICML*, 2007.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. *ICML*, 2008.

Wainwright, M., Ravikumar, P., and Lafferty, J. High dimensional graphical model selection using ℓ_1-regularized logistic regression. *NIPS*, 2006.

Yang, E. and Ravikumar, P. On the use of variational inference for learning discrete graphical models. *ICML*, 2011.

Younes, L. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré*, 1988.

A. Notation

We use the notation in Table 3.

Table 3. Notation used in this paper.

Notation	Description
$\|\mathbf{c}\|_1$	ℓ_1-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sum_n \|c_n\|$
$\|\mathbf{c}\|_\infty$	ℓ_∞-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\max_n \|c_n\|$
$\|\mathbf{c}\|_2$	Euclidean norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sqrt{\sum_n c_n^2}$
$\|\mathbf{A}\|_1$	ℓ_1-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} \|a_{mn}\|$
$\|\mathbf{A}\|_\infty$	ℓ_∞-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\max_{mn} \|a_{mn}\|$
$\|\mathbf{A}\|_\mathfrak{F}$	Frobenius norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sqrt{\sum_{mn} a_{mn}^2}$
$\langle \mathbf{A}, \mathbf{B} \rangle$	scalar product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} a_{mn} b_{mn}$
$\mathbf{diag}(\mathbf{A}) \in \mathbb{R}^N$	vector with diagonal elements of $\mathbf{A} \in \mathbb{R}^{N\times N}$
$\partial f / \partial \mathbf{c}$	gradient of f with respect to $\mathbf{c} \in \mathbb{R}^N$, i.e. $\partial f / \partial \mathbf{c} \in \mathbb{R}^N$
$\partial f / \partial \mathbf{A}$	gradient of f with respect to $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\partial f / \partial \mathbf{A} \in \mathbb{R}^{M\times N}$

B. Technical Lemma

The following technical lemma generalizes Lemma 1 of (Duchi & Singer, 2009) in which we assume a sequence of deterministic errors.

Lemma 10. *For a sequence of deterministic errors $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}$ and non-increasing step sizes η_k, the objective function evaluated at each iteration is bounded as follows:*

$$\eta_k(\mathcal{L}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^*)) + \eta_{k+1}(\mathcal{R}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{R}(\boldsymbol{\theta}^*)) \leq \tfrac{1}{2}\begin{pmatrix} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2 - \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|_2^2 \\ +4D\eta_k\|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2 \end{pmatrix} \tag{16}$$

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

As noted in (Duchi & Singer, 2009), eq.(8) can be written as a single step:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_\mathcal{R}^{(k+1)} \tag{17}$$

where $\mathbf{g}_\mathcal{R}^{(k+1)} \in \frac{\partial \mathcal{R}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^{(k+1)})$. This follows from the fact that $\boldsymbol{\theta}^{(k+1)}$ minimizes Step ii of eq.(8), if and only if $\mathbf{0}$ belongs to the subdifferential set of the non-smooth objective function evaluated at $\boldsymbol{\theta}^{(k+1)}$.

By eq.(17), $a^{(k+1)} = \|\boldsymbol{\theta}^{(k)} - \eta_k(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_\mathcal{R}^{(k+1)} - \boldsymbol{\theta}^*\|_2^2 = a^{(k)} + 2\eta_k F_1 + 2\eta_{k+1}F_2 + 2\eta_{k+1}F_3 + 2\eta_k F_4 + F_5$ for $F_1 \equiv -\langle \mathbf{g}_\mathcal{L}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^* \rangle$, $F_2 \equiv -\langle \mathbf{g}_\mathcal{R}^{(k+1)}, \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^* \rangle$, $F_3 \equiv \langle \mathbf{g}_\mathcal{R}^{(k+1)}, \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} \rangle$, $F_4 \equiv -\langle \boldsymbol{\xi}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^* \rangle$ and $F_5 \equiv \|\eta_k(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}) + \eta_{k+1}\mathbf{g}_\mathcal{R}^{(k+1)}\|_2^2$.

By the definition of subgradients of convex functions, $F_1 \leq \mathcal{L}^* - \mathcal{L}^{(k)}$ and $F_2 \leq \mathcal{R}^* - \mathcal{R}^{(k+1)}$.

By eq.(17), the Cauchy-Schwarz inequality and Assumption 5, $F_3 = \langle \mathbf{g}_\mathcal{R}^{(k+1)}, -\eta_k(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}) - \eta_{k+1}\mathbf{g}_\mathcal{R}^{(k+1)} \rangle \leq \|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2 \|\eta_k(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}) + \eta_{k+1}\mathbf{g}_\mathcal{R}^{(k+1)}\|_2 \leq \|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2(\eta_k\|\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}\|_2 + \eta_{k+1}\|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2) \leq (\eta_k + \eta_{k+1})G^2$.

By the Cauchy-Schwarz inequality, $F_4 \leq \|\boldsymbol{\xi}^{(k)}\|_2\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2 \leq D\|\boldsymbol{\xi}^{(k)}\|_2$, since by assumption $(\forall k)\ a^{(k)} \leq D^2$.

By the Cauchy-Schwarz inequality and Assumption 5, $F_5 \leq \eta_k^2\|(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)})\|_2^2 + 2\eta_k\eta_{k+1}\langle \mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}, \mathbf{g}_\mathcal{R}^{(k+1)} \rangle + \eta_{k+1}^2\|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2^2 \leq \eta_k^2\|(\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)})\|_2^2 + 2\eta_k\eta_{k+1}\|\mathbf{g}_\mathcal{L}^{(k)} + \boldsymbol{\xi}^{(k)}\|_2\|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2 + \eta_{k+1}^2\|\mathbf{g}_\mathcal{R}^{(k+1)}\|_2^2 \leq (\eta_k^2 + 2\eta_k\eta_{k+1} + \eta_{k+1}^2)G^2$.

Putting everything together, $a^{(k+1)} \leq a^{(k)} + 2\eta_k(\mathcal{L}^* - \mathcal{L}^{(k)}) + 2\eta_{k+1}(\mathcal{R}^* - \mathcal{R}^{(k+1)}) + 2\eta_k D\|\boldsymbol{\xi}^{(k)}\|_2 + (\eta_k^2 + 4\eta_k\eta_{k+1} + 3\eta_{k+1}^2)G^2$. Finally, since $\eta_{k+1} \leq \eta_k \Rightarrow (\eta_k^2 + 4\eta_k\eta_{k+1} + 3\eta_{k+1}^2)G^2 \leq 8\eta_k^2 G^2$. □

C. Detailed Proofs

In this section, we show the detailed proofs of lemmas and theorems for which we provide only proof sketches.

C.1. Proof of Theorem 2

Proof. For proving Claim i, note that for Ising models (and in general for any discrete probability distribution) the negative log-likelihood in eq.(2) is non-negative, i.e. $(\forall \mathbf{x})\ p_{\boldsymbol{\theta}}(\mathbf{x}) \in [0;1] \Rightarrow (\forall \mathbf{x})\ \log p_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0 \Rightarrow \mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{T}\sum_t \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}) \geq 0$. Given that $(\mathbf{W}^*, \mathbf{b}^*)$ is the optimal solution, $N\log 2 = \mathcal{L}(\mathbf{0}, \mathbf{0}) + \mathcal{R}(\mathbf{0}) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) + \mathcal{R}(\mathbf{W}^*) \geq \mathcal{R}(\mathbf{W}^*) = \rho\|\mathbf{W}^*\|_1$, and we prove our claim.

For proving Claim ii, note that the regularizer $\mathcal{R}(\mathbf{W})$ is non-negative, therefore $N\log 2 = \mathcal{L}(\mathbf{0}, \mathbf{0}) + \mathcal{R}(\mathbf{0}) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) + \mathcal{R}(\mathbf{W}^*) \geq \mathcal{L}(\mathbf{W}^*, \mathbf{b}^*) \geq \log(\sum_\mathbf{x} e^{-\|\mathbf{W}^*\|_1 + \mathbf{b}^{*\mathrm{T}}\mathbf{x}}) - \|\widehat{\boldsymbol{\Sigma}}\|_\infty\|\mathbf{W}^*\|_1 - \widehat{\boldsymbol{\mu}}^\mathrm{T}\mathbf{b}^* = -\|\mathbf{W}^*\|_1 + \log(\sum_\mathbf{x} e^{\mathbf{b}^{*\mathrm{T}}\mathbf{x}}) - \|\widehat{\boldsymbol{\Sigma}}\|_\infty\|\mathbf{W}^*\|_1 - \widehat{\boldsymbol{\mu}}^\mathrm{T}\mathbf{b}^* = \sum_n \log(e^{b_n^*} + e^{-b_n^*}) - \widehat{\boldsymbol{\mu}}^\mathrm{T}\mathbf{b}^* - (1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1 \geq \|\mathbf{b}^*\|_1 - \widehat{\boldsymbol{\mu}}^\mathrm{T}\mathbf{b}^* - (1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1 \geq (1 - \|\widehat{\boldsymbol{\mu}}\|_\infty)\|\mathbf{b}^*\|_1 - (1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1$. Recall that by Assumption 1, $\|\widehat{\boldsymbol{\mu}}\|_\infty < 1$. Therefore, $\|\mathbf{b}^*\|_1 \leq (N\log 2 + (1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)\|\mathbf{W}^*\|_1)/(1 - \|\widehat{\boldsymbol{\mu}}\|_\infty)$ and by using Claim i we prove our claim.

Claim iii follows from Claims i and ii and the fact that $\|\boldsymbol{\theta}^*\|_2^2 = \|\mathbf{W}^*\|_{\mathfrak{F}}^2 + \|\mathbf{b}^*\|_2^2 \leq \|\mathbf{W}^*\|_1^2 + \|\mathbf{b}^*\|_1^2$. □

C.2. Proof of Lemma 3

Proof. For proving Claim i, note that the terms $\partial \log \mathcal{Z}/\partial \mathbf{W}$ and $\partial \log \mathcal{Z}/\partial \mathbf{b}$ in eq.(4) are the second and first-order moment of binary variables in $\{-1, +1\}$.

Proving Claims ii and iii is straightforward from applying the above claims in eq.(4).

For proving Claim iv, recall that the subgradient $\partial \mathcal{R}/\partial \mathbf{W} = \{\mathbf{G} \mid \|\mathbf{G}\|_\infty \leq \rho \wedge \langle \mathbf{G}, \mathbf{W}\rangle = \|\mathbf{W}\|_1\}$. Therefore, $(\forall \mathbf{G} \in \partial \mathcal{R}/\partial \mathbf{W})\ \|\mathbf{G}\|_\infty \leq \rho$.

Claim v follows from Claims ii to iv and the fact that $\|\partial \mathcal{L}/\partial \boldsymbol{\theta}\|_2^2 = \|\partial \mathcal{L}/\partial \mathbf{W}\|_{\mathfrak{F}}^2 + \|\partial \mathcal{L}/\partial \mathbf{b}\|_2^2$. Furthermore, for the first term $\|\partial \mathcal{L}/\partial \mathbf{W}\|_{\mathfrak{F}} \leq N\|\partial \mathcal{L}/\partial \mathbf{W}\|_\infty \leq N(1 + \|\widehat{\boldsymbol{\Sigma}}\|_\infty)$, and for the second term $\|\partial \mathcal{L}/\partial \mathbf{b}\|_2 \leq \sqrt{N}\|\partial \mathcal{L}/\partial \mathbf{b}\|_\infty \leq \sqrt{N}(1 + \|\widehat{\boldsymbol{\mu}}\|_\infty)$. Similarly, $\|\partial \mathcal{R}/\partial \boldsymbol{\theta}\|_2 = \|\partial \mathcal{R}/\partial \mathbf{W}\|_{\mathfrak{F}} \leq N\|\partial \mathcal{R}/\partial \mathbf{W}\|_\infty \leq N\rho$. □

C.3. Proof of Theorem 6

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k \eta_k(\mathcal{L}^{(k)} + \mathcal{R}^{(k)})/\sum_k \eta_k$. Therefore $\mathcal{L}(\overline{\boldsymbol{\theta}}) - \mathcal{L}^* + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{R}^* \leq \pi_\eta(K) \leq (\eta_1\mathcal{R}^{(1)} + \sum_k (\eta_k(\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*)))/\sum_k \eta_k \equiv F$, and since $\boldsymbol{\theta}^{(1)} = \mathbf{0} \Rightarrow \mathcal{R}^{(1)} = 0$.

By Lemma 10 we know that $\eta_k(\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \frac{1}{2}(a^{(k)} - a^{(k+1)} + 4D\eta_k\|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2) \Rightarrow (\sum_k \eta_k)F \leq \frac{1}{2}\sum_k (a^{(k)} - a^{(k+1)}) + 2D(\sum_k \eta_k\|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2)G^2 \leq \frac{a^{(1)}}{2} + 2D(\sum_k \eta_k\|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2)G^2$.

Since by assumption $(\forall k)\ a^{(k)} \leq D^2 \Rightarrow (\sum_k \eta_k)F \leq \frac{D^2}{2} + 2D(\sum_k \eta_k\|\boldsymbol{\xi}^{(k)}\|_2) + 4(\sum_k \eta_k^2)G^2$. Finally, by replacing $\eta_k = \frac{\beta}{Gk^r}$, we prove our claim. □

C.4. Proof of Theorem 7

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $a^{(k)} \equiv \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2^2$.

By Jensen's inequality $\mathcal{L}(\overline{\boldsymbol{\theta}}) + \mathcal{R}(\overline{\boldsymbol{\theta}}) \leq \sum_k (\mathcal{L}^{(k)} + \mathcal{R}^{(k)})/K$. Therefore $\mathcal{L}(\overline{\boldsymbol{\theta}}) - \mathcal{L}^* + \mathcal{R}(\overline{\boldsymbol{\theta}}) - \mathcal{R}^* \leq \pi(K) \leq (\mathcal{R}^{(1)} + \sum_k (\mathcal{L}^{(k)} - \mathcal{L}^* + \mathcal{R}^{(k+1)} - \mathcal{R}^*))/K \equiv F$, and since $\boldsymbol{\theta}^{(1)} = \mathbf{0} \Rightarrow \mathcal{R}^{(1)} = 0$.

For using Lemma 10, note that since $\eta_{k+1} \leq \eta_k \Rightarrow \eta_{k+1}(\mathcal{L}^{(k)} - \mathcal{L}^* + \mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \eta_k(\mathcal{L}^{(k)} - \mathcal{L}^*) + \eta_{k+1}(\mathcal{R}^{(k+1)} - \mathcal{R}^*) \leq \frac{1}{2}(a^{(k)} - a^{(k+1)} + 4D\eta_k\|\boldsymbol{\xi}^{(k)}\|_2 + 8\eta_k^2 G^2)$. Furthermore, since $\frac{\eta_k}{\eta_{k+1}} \leq 2^r \Rightarrow KF \leq \frac{1}{2}\sum_k \frac{a^{(k)} - a^{(k+1)}}{\eta_{k+1}} + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2 \leq \frac{a^{(1)}}{2\eta_2} + \frac{1}{2}\sum_{k=2}^K \left(\frac{a^{(k)}}{\eta_{k+1}} - \frac{a^{(k)}}{\eta_k}\right) + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2$.

Since by assumption $(\forall k)\ a^{(k)} \leq D^2 \Rightarrow KF \leq \frac{D^2}{2}\left(\frac{1}{\eta_2} + \sum_{k=2}^K \left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k}\right)\right) + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2 \leq \frac{D^2}{2\eta_{K+1}} + 2^{1+r}D(\sum_k \|\boldsymbol{\xi}^{(k)}\|_2) + 2^{2+r}(\sum_k \eta_k)G^2$. Finally, by replacing $\eta_k = \frac{\beta}{Gk^r}$, we prove our claim. □

C.5. Proof of Theorem 8

Proof. Let $\mathcal{L}^{(k)} \equiv \mathcal{L}(\boldsymbol{\theta}^{(k)})$, $\mathcal{R}^{(k)} \equiv \mathcal{R}(\boldsymbol{\theta}^{(k)})$, $\mathcal{L}^* \equiv \mathcal{L}(\boldsymbol{\theta}^*)$, $\mathcal{R}^* \equiv \mathcal{R}(\boldsymbol{\theta}^*)$ and $\mathcal{U}$ the uniform distribution for $k \in \{1, \ldots, K\}$.

By Markov's inequality, for $a^{(k)} = \mathcal{L}^{(k)} + \mathcal{R}^{(k)} - \mathcal{L}^* - \mathcal{R}^* \geq 0$, we have $\mathbb{P}_{\mathcal{U}}[a^{(k)} \geq c] \leq \frac{\mathbb{E}_{\mathcal{U}}[a^{(k)}]}{c}$. Note that $\mathbb{E}_{\mathcal{U}}[a^{(k)}] = \frac{1}{K}\sum_k (\mathcal{L}^{(k)} + \mathcal{R}^{(k)}) - \mathcal{L}^* - \mathcal{R}^* = \pi(K)$. By Theorem 7, we know that $\pi(K) \leq \frac{D^2 G(K+1)^r}{2\beta K} + 2^{1+r}DA_{\boldsymbol{\gamma},\boldsymbol{\xi}} + \frac{2^{2+r}\beta G H_{r,K}}{K} \equiv F$, therefore $\mathbb{P}_{\mathcal{U}}[a^{(k)} \geq c] \leq \frac{F}{c}$. For $c = \frac{F}{\varepsilon} \Rightarrow \mathbb{P}_{\mathcal{U}}[a^{(k)} \geq \frac{F}{\varepsilon}] \leq \varepsilon$. □

C.6. Proof of Theorem 9

Proof. Let $\mathcal{Q}_k$ be the distribution of the error for the k-th sampler, the joint distribution $\mathcal{Q} \equiv \{\mathcal{Q}_1, \ldots, \mathcal{Q}_K\}$, $\mathcal{T}$ be the joint distribution of $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\}$, the first-order moment $\phi_k \equiv \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]$ and the second-order moment $\nu_k^2 \equiv \mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]$.

By Lemma 3 we know that $\|\partial \log \mathcal{Z}/\partial \boldsymbol{\theta}\|_\infty \leq 1$. By Assumption 5, for any sampler we have $\|\partial \log \mathcal{Z}/\partial \boldsymbol{\theta} + \boldsymbol{\xi}^{(k)}\|_\infty \leq 1$ and therefore $\|\boldsymbol{\xi}^{(k)}\|_\infty \leq 2$ in the worst case. Therefore $\|\boldsymbol{\xi}^{(k)}\|_2 \leq \sqrt{M}\|\boldsymbol{\xi}^{(k)}\|_\infty \leq 2\sqrt{M}$.

Given that the error is bounded, we have $\mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2] \leq 2\sqrt{M}$. By using the bounds in Definition 4, the bias is at most $\phi_k \leq \min(2\sqrt{M}, \frac{B}{S_k})$.

Similarly, we have $\mathbb{V}\mathrm{ar}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2] = \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2^2] - \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2]^2 \leq \mathbb{E}_{\mathcal{Q}}[\|\boldsymbol{\xi}^{(k)}\|_2^2] \leq 4M$. By using the bounds in Definition 4, the variance is at most $\nu_k^2 \leq \min(4M, \frac{V}{S_k})$.

Consider the variable $z_k = K\gamma_k\|\boldsymbol{\xi}^{(k)}\|_2$. Note that the mean $\widehat{z} = \frac{1}{K}\sum_k z_k = \sum_k \gamma_k\|\boldsymbol{\xi}^{(k)}\|_2 = A_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is the expression we want to upper-bound. The expected value $\overline{\phi} = \mathbb{E}_{\mathcal{Q}}[\widehat{z}] = \sum_k \gamma_k\phi_k \leq \min(2\sqrt{M}, B\sum_k \frac{\gamma_k}{S_k}) \equiv \lambda_1$.

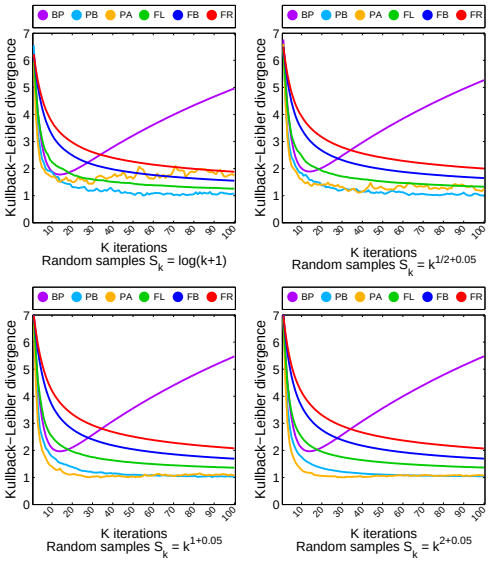

Figure 2. Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "zero-field" regime and "Gibbs sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.

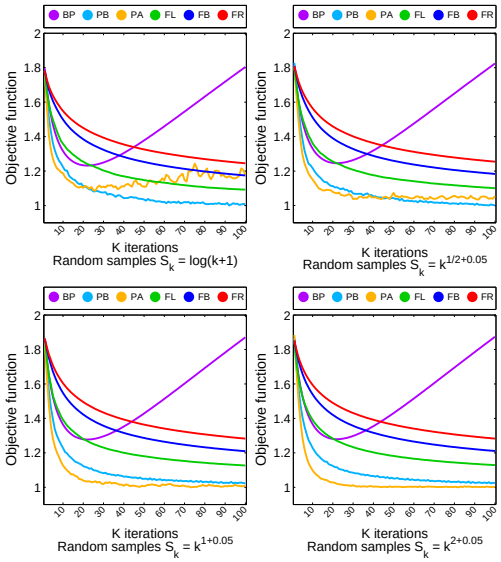

Figure 3. Objective function for different settings of increasing number of random samples for the "non-zero field" regime and "Gibbs sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

The average variance $\sigma^2 = \frac{1}{K}\sum_k \mathbb{V}\text{ar}_{\mathcal{Q}}[z_k] = K\sum_k \gamma_k^2\nu_k^2 \leq K\min(4M\sum_k \gamma_k^2, V\sum_k \frac{\gamma_k^2}{S_k}) \leq K\min(4M, V\sum_k \frac{\gamma_k^2}{S_k}) \equiv K\lambda_2$.

Our goal is to find an upper bound for $F_1 \equiv \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \lambda_1 + \epsilon]$. By the definition of marginal distribution $F_1 = \int_{\boldsymbol{\Theta}} \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \lambda_1 + \epsilon \mid \boldsymbol{\Theta}] p_{\mathcal{T}}(\boldsymbol{\Theta}) \leq \int_{\boldsymbol{\Theta}} \mathbb{P}_{\mathcal{Q}}[\widehat{z} \geq \overline{\phi} + \epsilon \mid \boldsymbol{\Theta}] p_{\mathcal{T}}(\boldsymbol{\Theta}) \equiv F_2$. By Bernstein inequality, $F_2 \leq \int_{\boldsymbol{\Theta}} e^{-\frac{K\epsilon^2}{2\sigma^2+4\sqrt{M}\epsilon/3}} p_{\mathcal{T}}(\boldsymbol{\Theta}) \leq \int_{\boldsymbol{\Theta}} e^{-\frac{K\epsilon^2}{2K\lambda_2+4\sqrt{M}\epsilon/3}} p_{\mathcal{T}}(\boldsymbol{\Theta}) = e^{-\frac{K\epsilon^2}{2K\lambda_2+4\sqrt{M}\epsilon/3}} \int_{\boldsymbol{\Theta}} p_{\mathcal{T}}(\boldsymbol{\Theta}) = e^{-\frac{K\epsilon^2}{2K\lambda_2+4\sqrt{M}\epsilon/3}} = \delta$. By solving for ϵ in the last equality, we prove our claim. □

D. Additional Experimental Results

First, we complement the results in Figure 1. We show the Kullback-Leibler divergence to the ground truth in Figure 2.

Note that we assumed a "zero field" regime for Figures 1 and 2 where $\mathbf{b}_g = \mathbf{0}$. We also report results in Figures 3 and 4 for the "non-zero field" regime where each entry of $\mathbf{b}_g$ is generated uniformly at random from $[-1; +1]$.

We also evaluate a "mean field sampler" by first finding the mean field distribution and then performing importance sampling with the mean field trial. We report results for the "zero field" regime in Figures 5

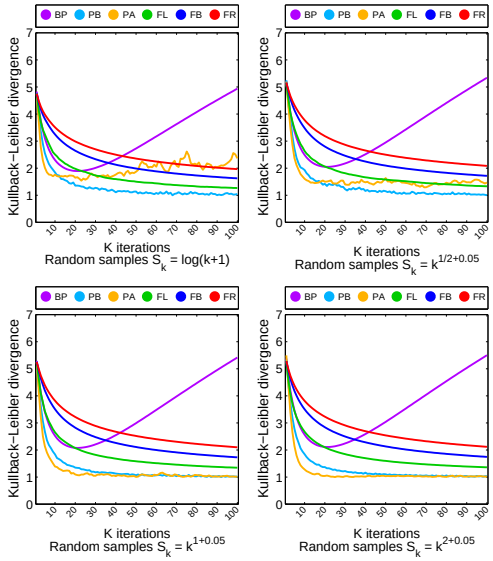

Figure 4. Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "non-zero field" regime and "Gibbs sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.

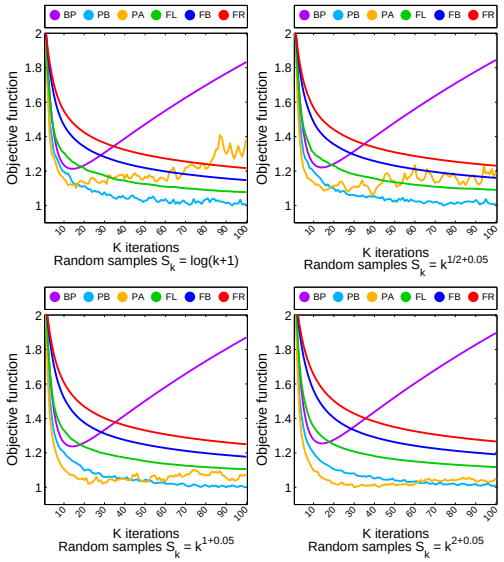

Figure 5. Objective function for different settings of increasing number of random samples for the "zero-field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

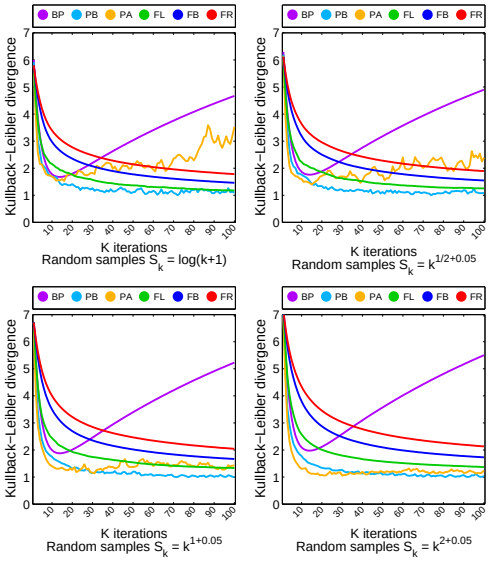

Figure 6. Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "zero-field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.

and 6, and for the "non-zero field" regime in Figures 7 and 8.

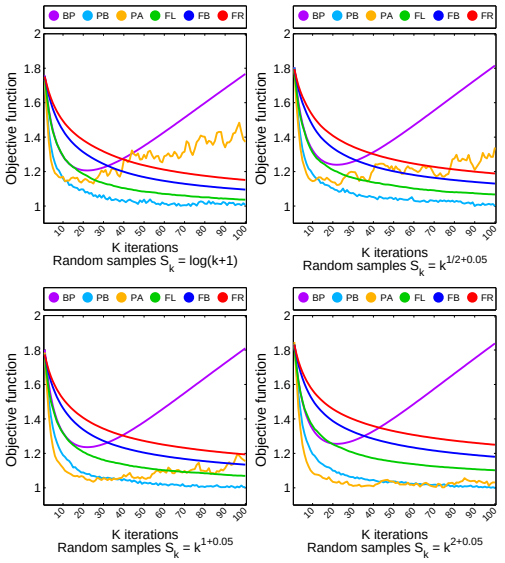

Figure 7. Objective function for different settings of increasing number of random samples for the "non-zero field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to converge, but they exhibit faster convergence. Belief propagation (BP) does not converge.

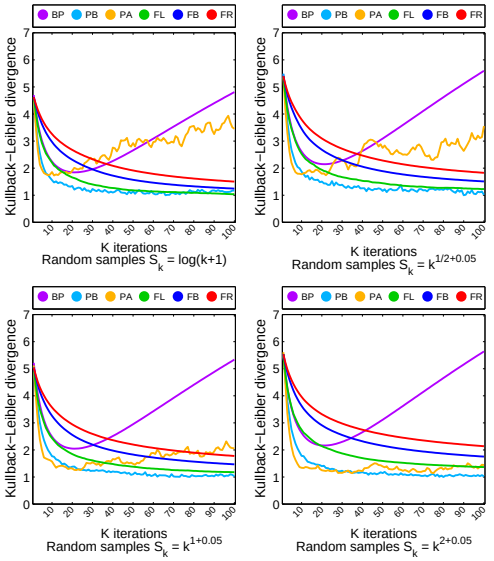

Figure 8. Kullback-Leibler divergence to the ground truth for different settings of increasing number of random samples for the "non-zero field" regime and "mean field sampler". Basic (PB) and accelerated (PA) are noisier and require more samples than last point (FL), basic (FB) and robust (FR) forward-backward splitting in order to generalize well, but they exhibit faster convergence. Belief propagation (BP) does not generalize well.