STOCHASTIC VARIATIONAL INFERENCE FOR LARGE SCALE MACHINE LEARNING

VARIATIONAL BAYES: APPROXIMATE POSTERIOR

$$p(\underset{\sim}{z}, \underset{\sim}{\theta} \mid \underset{\approx}{x}) = \frac{\prod_{i=1}^{M} p(\underset{\sim}{z}_i, \underset{\sim}{x}_i \mid \underset{\sim}{\theta}) p(\underset{\sim}{\theta})}{p(\underset{\approx}{x})}$$

WITH SIMPLER DISTRIBUTION $q_\lambda(\underset{\sim}{z}, \underset{\sim}{\theta})$. PARAMETERIZED BY A SET OF PARAMETERS λ.

VB MAXIMIZES THE EVIDENCE LOWER BOUND OR ELBO

$$\log p(\underset{\approx}{x}) = \log \mathbb{E}_{q_\lambda}\left[\frac{p(\underset{\approx}{x}, \underset{\sim}{z}, \underset{\sim}{\theta})}{q_\lambda(\underset{\sim}{z}, \underset{\sim}{\theta})}\right] \geq$$

$$\geq \underbrace{\mathbb{E}_{q_\lambda}\left[\log p(\underset{\approx}{x}, \underset{\sim}{z}, \underset{\sim}{\theta})\right] - \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\underset{\sim}{z}, \underset{\sim}{\theta})\right]}_{\mathcal{L}(\lambda)}$$

LET'S ASSUME THAT

$$\mathcal{L}(\lambda) = \frac{1}{M} \sum_{i=1}^{M} f(\underset{\sim}{x}_i, \lambda_t)$$

AND THAT λ IS A VECTOR. WHAT TECHNIQUES CAN WE USE TO OPTIMIZE $\mathcal{L}(\underset{\sim}{\lambda})$?

GRADIENT DESCENT (GD).

EACH ITERATION UPDATES λ USING

$$\lambda_{t+1} = \lambda_t + \gamma \frac{1}{M} \sum_{i=1}^{M} \nabla_\lambda f(\underset{\sim}{x}_i, \lambda_t)$$

UNDER SUFFICIENT REGULARITY ASSUMPTIONS, WHEN γ IS SMALL ENOUGH, GD HAS LINEAR CONVERGENCE: $-\log \rho \sim t$ WHERE ρ IS THE RESIDUAL ERROR: $\rho = |\lambda_t - \lambda^*|$

SECOND ORDER GRADIENT DESCENT (2GD)

EACH ITERATION UPDATES λ USING

$$\lambda_{t+1} = \lambda_t + \Gamma_t \frac{1}{M} \sum_{i=1}^{M} \nabla_\lambda f(\underset{\sim}{x}_i, \lambda_t)$$

Γ_t APPROACHES THE INVERSE OF THE HESSIAN AT THE OPTIMUM.

UNDER SUF. REGULARITY ASSUMPTIONS 2GD HAS QUADRATIC CONVERGENCE: $\log - \log \rho \sim t$

STOCHASTIC GRADIENT DESCENT.

$$\lambda_{t+1} = \lambda_t + \gamma_t \nabla_\lambda f(\underset{\sim}{x}_{i_t}, \lambda_t)$$

WHERE i_t IS SAMPLED UNIFORMLY IN $\{1, \ldots, M\}$.

CONVERGENCE REQUIRES

$$\sum_t \gamma_t^2 < \infty \qquad \sum_t \gamma_t = \infty$$

WE CAN CHOOSE $\gamma_t = (t+\tau)^{-\kappa}$ WITH $\kappa \in (0.5, 1]$. UNDER SOME CONDITIONS (→ REGULARITY) THE BEST CONVERGENCE SPEED IS ACHIEVED WHEN $\gamma_t \sim t^{-1}$ IN THIS CASE $\mathbb{E}[\rho] \sim t^{-1}$.

ASYMPTOTIC BEHAVIOR OF EACH METHOD:

	GD	ZGD	SGD
TIME PER ITERATION	M	M	1
ITERATIONS TO ACCURACY ρ	$\lg \frac{1}{\rho}$	$\lg \lg 1/\rho$	$1/\rho$
TIME TO ACCURACY ρ	$M \lg 1/\rho$	$M \lg \lg 1/\rho$	$1/\rho$

DOES NOT DEPEND ON M!!

LET'S APPLY SGD TO OUR VARIATIONAL PROBLEM.

WE ASSUME CONJUGATE EXPONENTIAL FAMILIES FOR OUR PRIOR AND LIKELIHOOD.

$$p(\underset{\sim}{\theta}) \propto \exp\{ \underset{\sim}{g}(\underset{\sim}{\theta})^T \underset{\sim}{\alpha} \}$$

$$p(\underset{\sim}{x}_i, \underset{\sim}{z}_i \mid \underset{\sim}{\theta}) \propto \exp\{ \underset{\sim}{g}(\underset{\sim}{\theta})^T \underset{\sim}{t}(x_i, z_i) \}$$

THE CONDITIONALS ARE IN THE SAME EXPONENTIAL FAMILY AS THE PRIOR.

$$p(\underset{\sim}{z}_i \mid \underset{\sim}{x}_i, \underset{\sim}{\theta}) \propto \exp\{ \underset{\sim}{\eta}(\underset{\sim}{x}_i, \underset{\sim}{\theta})^T \underset{\sim}{t}(z_i) \}$$

$$p(\underset{\sim}{\theta} \mid \underset{\sim}{x}, \underset{\sim}{z}) \propto \exp\{ \underset{\sim}{g}(\underset{\sim}{\theta})^T \left(\underset{\sim}{\alpha} + \sum_{i=1}^{M} \underset{\sim}{t}(\underset{\sim}{x}_i, \underset{\sim}{z}_i) \right) \}$$

WE CHOOSE FOR q_λ A FACTORIZED FORM (MEAN-FIELD).

$$q_\lambda(\underset{\sim}{\theta}, \underset{\sim}{z}) = q_\lambda(\underset{\sim}{\theta}) \cdot \prod q_\lambda(z_i)$$

THE OPTIMAL FACTORS FOR q_λ ARE IN THE SAME FAMILY AS THE COMPLETE CONDITIONALS WHY?

$$q(z_i) = \exp\{ \underset{\sim}{\beta}_i^T \underset{\sim}{t}(z_i) - z(\beta_i) \}$$

$$q(\underset{\sim}{\theta} \mid \underset{\sim}{\gamma}) = \exp\{ \underset{\sim}{g}(\theta)^T \underset{\sim}{\gamma} - z(\underset{\sim}{\gamma}) \}$$

$\lambda = \{ \underset{\sim}{\theta}, \beta_1, \ldots, \beta_M \}$

AS A FUNCTION OF $\underset{\sim}{\gamma}$ THE ELBO IS

$$\mathcal{L}(\underset{\sim}{\gamma}) = \mathbb{E}_q[\log p(\underset{\sim}{\theta}, \underset{\sim}{x}, \underset{\sim}{z})] - \mathbb{E}_q[\log q(\underset{\sim}{\theta})] + \text{const.}$$

$$\mathcal{L}(\underset{\sim}{\gamma}) = \mathbb{E}_q\left[\underset{\sim}{\alpha} + \sum_{i=1}^{M} \underset{\sim}{t}(x_i, z_i)\right]\left[\nabla_{\underset{\sim}{\gamma}} z(\underset{\sim}{\gamma})\right] -$$

$$- \underset{\sim}{\gamma}^T \nabla_{\underset{\sim}{\gamma}} z(\underset{\sim}{\gamma}) + z(\underset{\sim}{\gamma}) + \text{CONST.}$$

$$\boxed{\nabla_{\underset{\sim}{\gamma}} z(\underset{\sim}{\gamma}) = \mathbb{E}_q[g(\underset{\sim}{\theta})]}$$

AND ITS GRADIENT IS

$$\nabla_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma}) = \nabla^2_{\underset{\sim}{\gamma}} z(\underset{\sim}{\gamma}) \Big\{ \mathbb{E}_q\left[\underset{\sim}{\alpha} + \sum_{i=1}^{M} \underset{\sim}{t}(x_i, z_i)\right]$$

$$- \underset{\sim}{\gamma} \Big\}$$

THE ELBO IS OPTIMIZED W.R.T. $\underset{\sim}{\gamma}$ BY SETTING $\underset{\sim}{\gamma} = \mathbb{E}_q\left[\underset{\sim}{\alpha} + \sum_{i=1}^{M} \underset{\sim}{t}(x_i, z_i)\right]$

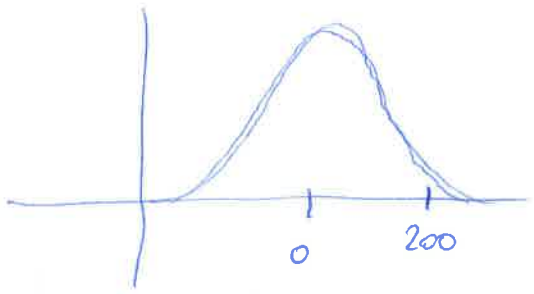
W.R.T. β_c BY SETTING $\underset{\sim}{\beta}_c = \mathbb{E}_q\left[\underset{\sim}{\eta}(x_i, \underset{\sim}{\theta})\right]$.

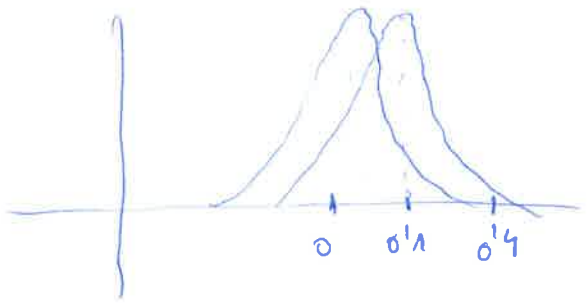
SLOW CONVERGENCE !!!

ALTERNATIVE : SGD. USE NATURAL GRADIENTS INSTEAD OF STANDARD GRADIENTS.

EUCLIDIAN DISTANCE BETWEEN $N(0, 10^4)$ AND $N(10, 10^4)$ IS 10.

DISTANCE BETWEEN $N(0, 10^{-2})$ AND $N(10^{-1}, 10^{-2})$ IS 10^{-1}





EUCLIDIANI DISTANCE BETWEEN $q(\underset{\sim}{\theta}|\underset{\sim}{\gamma})$ AND $q(\underset{\sim}{\theta}|\underset{\sim}{\gamma}')$ IS $\| \underset{\sim}{\gamma} - \underset{\sim}{\gamma}' \|_1$

USE KL-DISTANCE : ~~KL(γ||γ')~~

↓

SYMETRIZED DIST $\left[q(\underset{\sim}{\theta}|\underset{\sim}{\gamma}), q(\underset{\sim}{\theta}|\underset{\sim}{\gamma}') \right]$:

$$= \mathbb{E}_{q_{\underset{\sim}{\gamma}}} \left[\log \frac{q(\underset{\sim}{\theta}|\underset{\sim}{\gamma})}{q(\underset{\sim}{\theta}|\underset{\sim}{\gamma}')} \right] + \mathbb{E}_{q_{\underset{\sim}{\gamma}'}} \left[\log \frac{q(\underset{\sim}{\theta}|\underset{\sim}{\gamma}')}{q(\underset{\sim}{\theta}|\underset{\sim}{\gamma})} \right]$$

THE DIRECTION OF THE NATURAL GRADIENT IS

$$\underset{\Delta \underset{\sim}{\gamma}}{\text{ARG MAX}} \; \mathcal{L}(\underset{\sim}{\gamma} + \Delta\underset{\sim}{\gamma}) \quad \text{s.t.} \quad KL^{SYM}(\underset{\sim}{\gamma}, \underset{\sim}{\gamma} + \Delta\underset{\sim}{\gamma}) < \epsilon$$

WHEN $\epsilon \to 0$.

THE NATURAL GRADIENT CAN BE SHOWN TO BE

$$\tilde{\nabla}^*_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma}) = \underset{\sim}{G}(\underset{\sim}{\gamma})^{-1} \nabla_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma})$$

RIEMANIAN METRIC TENSOR

$$\Delta\underset{\sim}{\gamma}^T \underset{\sim}{G}(\underset{\sim}{\gamma}) \Delta\underset{\sim}{\gamma} =$$

$$KL^{SYM}(\underset{\sim}{\gamma}, \underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma}) = \cancel{E_{q_{\underset{\sim}{\gamma}}}[\,]}$$

$$= \nabla z(\underset{\sim}{\gamma})^T \underset{\sim}{\gamma} - z(\underset{\sim}{\gamma}) - \nabla z(\underset{\sim}{\gamma})^T(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma}) + z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma})$$

$$+ \nabla z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma})^T(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma}) - z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma}) -$$

$$- \nabla z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma})^T\underset{\sim}{\gamma} + z(\underset{\sim}{\gamma})$$

$$= -\nabla z(\underset{\sim}{\gamma})^T \Delta\underset{\sim}{\gamma} + \nabla z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma})^T \Delta\underset{\sim}{\gamma}$$

$$= \left[\nabla z(\underset{\sim}{\gamma}+\Delta\underset{\sim}{\gamma})^T - \nabla z(\underset{\sim}{\gamma})^T\right]\Delta\underset{\sim}{\gamma}$$

$$= \cancel{\Delta\underset{\sim}{\gamma}^T \nabla^2 z(\underset{\sim}{\gamma})\,\Delta\underset{\sim}{\gamma}}$$

$$= \cancel{\left[\nabla z(\underset{\sim}{\gamma})^T + \Delta\gamma\nabla^2 z(\underset{\sim}{\gamma}) - \nabla z(\underset{\sim}{\gamma})^T\right]}\Delta\underset{\sim}{\gamma} =$$

$$= \Delta\underset{\sim}{\gamma}\,\nabla^2 z(\underset{\sim}{\gamma})\,\Delta\underset{\sim}{\gamma}$$

THEREFORE $\underset{\sim}{G}^{*}(\underset{\sim}{\gamma}) = \nabla^2 z(\underset{\sim}{\gamma})$.

AND $$\nabla^{*}_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma}) = E_q\left[\underset{\sim}{\alpha} + \sum_{i=1}^{M} \underset{\sim}{t}(x_i, z_i)\right] - \underset{\sim}{\gamma}$$

NOISY ESTIMATE OF THE NATURAL GRADIENT.

$$\nabla^{*\,NOISY}_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma}) = \alpha + \frac{1}{p(\underset{\sim}{\gamma})} E_q\left[\underset{\sim}{t}(x_i, z_i)\right] - \underset{\sim}{\gamma}$$

WHERE $p(k)$ IS THE PROB. OF SUBSAMPLING THE k-TH DATA POINT.

THE SVI UPDATE FOR $\underset{\sim}{\gamma}$ IS

$$\underset{\sim}{\gamma}^{new} = \underset{\sim}{\gamma}^{old} + \rho_t \tilde{\nabla}^{*noisy}_{\underset{\sim}{\gamma}} \mathcal{L}(\underset{\sim}{\gamma}) =$$

$$= (1-\rho_t)\underset{\sim}{\gamma}^{old} + \rho_t \underset{\sim}{\gamma}^{noisy}$$

WHERE $\underset{\sim}{\gamma}^{noisy} = \underset{\sim}{\alpha} + \frac{1}{p(k)} \mathbb{E}_q \left[\underset{\sim}{t}(\underset{\sim}{x}_k, z_k) \right]$

OPTIMAL PARAMETER IF THE DATA INCLUDES ONLY DATA k REPLICATED $\frac{1}{p(k)}$ TIMES.

WE CAN USE MINIBATCHES OF SIZE S TO REDUCE VARIANCE. IN THIS CASE

$$\underset{\sim}{\gamma}^{new} = (1-\rho_t)\underset{\sim}{\gamma}^{old} + \frac{1}{S} \sum_{i=1}^{S} \underset{\sim}{\gamma}^{noisy}(i)$$

SVI PSEUDO CODE:

REPEAT

SAMPLE $\underset{\sim}{x}_k$ WITH PROB. $p(k)$.

OPTIMALLY UPDATE $\underset{\sim}{\phi}_k = \mathbb{E}_q \left[\underset{\sim}{\eta}(\underset{\sim}{x}_k, \underset{\sim}{\theta}) \right]$

COMPUTE NOISY OPTIMAL $\underset{\sim}{\gamma}$

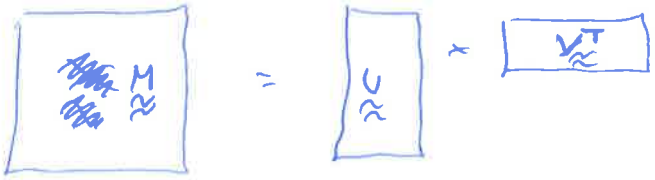
$$\tilde{\gamma}_{noisy} = \underset{\sim}{\alpha} + \frac{1}{p(k)} \mathbb{E}_q \left[\underset{\sim}{t}(\underset{\sim}{x}_k, z_k) \right]$$

UPDATE λ USING $\lambda^{new} = (1-\rho_t)\lambda^{old} + \rho_t \lambda^{noisy}$

UNTIL FOREVER.

$$\rho_t = (t+\tau)^{-\kappa} \quad \kappa \in (0.5, 1], \ \tau \geq 0.$$

SVI IN PROB. BAYESIAN MATRIX FACTORIZATION.



M IS AN $n \times d$ MATRIX

U AND V ARE $n \times k$ AND $d \times k$ MATRICES

$$p(M \mid U, V) = \left[\prod_{i=1}^{n}\prod_{j=1}^{d} N(m_{ij} \mid u_i v_j^T, \sigma^2)\right] \cdot \left[\prod_{i=1}^{n}\prod_{l=1}^{k} N(u_{il} \mid 0, \tau^2)\right] \left[\prod_{j=1}^{d}\prod_{l=1}^{k} N(v_{jl} \mid 0, \rho^2)\right]$$

THE COMPLETE CONDITIONALS ARE GAUSSIANS.

THE LIKELIHOOD AND THE PRIOR ARE ~~CONJUGATE~~ CONJUGATE.

$$Q(U,V) = \left[\prod_{i=1}^{n} N(u_i \mid m_i^u, V_i^u)\right] \left[\prod_{j=1}^{d} N(v_j \mid m_j^v, V_j^v)\right].$$

WE WORK WITH THE NATURAL PARAMETERS

$\underset{\sim}{\eta}^{u}_{2,i} = [\underset{\sim}{V}^{u}_{i}]^{-1}$ AND $\underset{\sim}{\eta}^{u}_{1,i} = [\underset{\sim}{V}^{u}_{i}]^{-1} \underset{\sim}{\mu}^{u}_{i}$

THE BATCH UPDATES ARE

$$\underset{\sim}{\eta}^{u}_{2,i} = \left[\sum_{j=1}^{d} \frac{\underset{\sim}{V}^{v}_{j} + \underset{\sim}{\mu}^{v}_{j} \underset{\sim}{\mu}^{v\,T}_{j}}{\sigma^2} \right] + \frac{1}{\tau^2} \underset{\sim}{I}$$

$$\underset{\sim}{\eta}^{u}_{1,i} = \sum_{j=1}^{d} \frac{m_{ij}\, \underset{\sim}{\mu}^{v}_{j}}{\sigma^2}$$

THE SVI UPDATES ARE:

$$\underset{\sim}{\eta}^{u\ new}_{2,i} = \underset{\sim}{\eta}^{u\ old}_{2,i} (1 - \rho_t) + \rho_t \underset{\sim}{\eta}^{u\ noisy\ minibatch}_{2,i}$$

$$\underset{\sim}{\eta}^{u\ new}_{1,i} = \underset{\sim}{\eta}^{u\ old}_{1,i} (1 - \rho_t) + \rho_t \underset{\sim}{\eta}^{u\ noisy\ minibatch}_{1,i}$$

WHERE

$$\underset{\sim}{\eta}^{u\ noisy\ minibatch}_{2,i} = \frac{d}{|I|} \left[\sum_{j \in I} \frac{\underset{\sim}{V}^{v}_{j} + \underset{\sim}{\mu}^{v}_{j} \underset{\sim}{\mu}^{v\,T}_{j}}{\sigma^2} \right] + \frac{1}{\tau^2} \underset{\sim}{I}$$

$$\underset{\sim}{\eta}^{u\ noisy\ minibatch}_{1,i} = \frac{d}{|I|} \left[\sum_{j \in I} \frac{m_{ij}\, \underset{\sim}{\mu}^{v}_{j}}{\sigma^2} \right]$$

WHERE I IS A RANDOM SET OF COLUMN INDEXES, SUCH THAT EACH ELEMENT OF I IS SAMPLED UNIFORMLY WITH REPLACEMENT FROM $\{1, \ldots, d\}$.

USEFULL REFERENCES:

VARIATIONAL BAYES:

1999 ATTIAS UAI.

2001 GHAHRAMANI, BEAL. ADVANCE MEAN FIELD METHOD — THEORY AND PRACTICE, CHAPTER GRAPHICAL MODELS AND VARIATIONAL METHODS.

2006 BISHOP PATTERN RECOGNITION AND MACHINE LEARNING.

STOCHASTIC GRADIENT DESCENT:

2010 BOTTOU LARGE SCALE MACHINE LEARNING WITH STOCHASTIC GRADIENT DESCENT.

1998 BOTTOU ON LINE LEARNING AND STOCHASTIC APPROXIMATIONS

1983 DENNIS SCHNABEL NUMERICAL METHODS FOR UNCONSTRAINED OPTIMIZATIONS AND NON-LINEAR EQUATIONS.

STOCHASTIC VARIATIONAL INFERENCE

2013 HOFFMAN, BLEI, CHONG, PAISLEY
STOCHASTIC VARIATIONAL INFERENCE

NATURAL GRADIENTS

1998 AMARI
NATURAL GRADIENT WORKS
EFFICIENTLY IN LEARNING